Genetic barriers more than environmental associations explain *Serratia marcescens* population structure

Check for updates

Lodovico Sterzi [1], Riccardo Nodari [1], Federico Di Marco [2], Maria Laura Ferrando[2], Francesca Saluzzo[2], Andrea Spitaleri[3], Hamed Allahverdi [1], Stella Papaleo [1], Simona Panelli [1], Sara Giordana Rimoldi[4], Gherard Batisti Biffignandi[5], Marta Corbella [6], Annalisa Cavallero[7], Paola Prati[8], Claudio Farina[9], Daniela Maria Cirillo [2], Gianvincenzo Zuccotti [1,10], Claudio Bandi [11] & Francesco Comandatore [1] ✉

Bacterial species often comprise well-separated lineages, likely emerged and maintained by genetic isolation and/or ecological divergence. How these two evolutionary actors interact in the shaping of bacterial population structure is currently not fully understood. In this study, we investigate the genetic and ecological drivers underlying the evolution of *Serratia marcescens*, an opportunistic pathogen with high genomic flexibility and able to colonise diverse environments. Comparative genomic analyses reveal a population structure composed of five deeply-demarcated genetic clusters with open pan-genome but limited inter-cluster gene flow, partially explained by Restriction-Modification (R-M) systems incompatibility. Furthermore, a large-scale research on hundred-thousands metagenomic datasets reveals only a partial habitat separation of the clusters. Globally, two clusters only show a separate gene composition coherent with ecological adaptations. These results suggest that genetic isolation has preceded ecological adaptations in the shaping of the species diversity, an evolutionary scenario coherent with the Evolutionary Extended Synthesis.

The evolutionary processes shaping the structure of bacterial populations have been deeply investigated and several speciation models have been proposed[1–3]. These models revolve mainly around the two most important mechanisms of genetic variation: mutation and recombination. In 2001, Cohan proposed the ecotype model of speciation[4], which focuses on the role of ecological divergence and selection. In absence of DNA exchange, bacterial lineages periodically accumulate mutations and diverge until one highly adapted lineage emerges and out-competes the other lineages, causing a clonal sweep phenomenon. Following this model, stable populations can only exist if they are ecologically diverse enough to avoid

competition. A more recent theoretical framework relies on barriers to recombination to explain the origin and maintenance of divergent sequence clusters[5–8], coherently with the Evolutionary Extended Synthesis view[9]. According to this view, the genetic cohesion is maintained by the persistent genetic exchange between the strains. A speciation event begins when a novel habitat-specific adaptive allele spreads within a subpopulation, conferring the ability to occupy a novel ecological niche. The ecological differentiation provides an initial barrier for recombination reducing the genetic exchange among the subpopulations. This process enhances the genetic divergence among these subpopulations, favouring the emergence of

[1]Department of Biomedical and Clinical Sciences, Pediatric Clinical Research Center "Romeo and Enrica Invernizzi", Università Di Milano, 20157 Milan, Italy. [2]Emerging Bacterial Pathogens Unit, Division of Immunology, Transplantation and Infectious Diseases, IRCCS San Raffaele Scientific Institute, Milan, Italy. [3]Neurogenomics Research Centre, Human Technopole, Milan, Italy. [4]Laboratorio di Microbiologia Clinica, Virologia e Diagnostica delle Bioemergenze, ASST Fatebenefratelli Sacco, Milan, Italy. [5]Department of Clinical-Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy. [6]Department of Microbiology & Virology, Fondazione IRCCS Policlinico San Matteo, Viale Camillo Golgi 19, 27100 Pavia, Italy. [7]Microbiology Unit, Fondazione IRCCS San Gerardo, Monza, Italy. [8]Istituto Zooprofilattico Sperimentale della Lombardia e dell'Emilia Romagna (IZSLER), Pavia, Italy. [9]Laboratorio di Microbiologia e Virologia, Azienda Socio-Sanitaria Territoriale (ASST) Papa Giovanni XXIII, Bergamo, Italy. [10]Department of Paediatrics, Children's Hospital "V. Buzzi", Milano, Italy. [11]Department of Biosciences and Pediatric Clinical Research Center "Romeo Ed Enrica Invernizzi", University of Milan, 20133 Milan, Italy. ✉e-mail: francesco.comandatore@unimi.it

genetic barriers and the formation of separate cohesive genotypic clusters. Indeed, recombination rates decrease drastically with sequence divergence[10–12]. This is due to the absence of stretches of identical nucleotides at one or both ends of the recombining DNA sequence, and to the incompatibility between Restriction-Modification (R-M) systems[6,13,14]. The R-M systems are the most widespread bacterial defence systems and rely on a straightforward, efficient mechanism to remove exogenous DNA: a methyltransferase methylates a specific sequence motif on the endogenous DNA and a cognate restriction endonuclease cleaves DNA when the motif is unmethylated. Thus, bacterial populations encoding for noncognate Restriction endonuclease-Methyltransferase (R-M) systems have fewer successful exchanges of genetic material.

Ultimately, what emerges is a picture in which bacterial differentiation must be viewed in light of two separate but not exclusive evolutionary drivers: ecology and genetic recombination. More wood has been added to the fire when the concept of pan-genome broke into bacterial population genomics. Often, the strains in a bacterial species share only a portion of their gene repertoire, while a consistent part of the genes are owned only by a few strains or lineages (accessory genes)[15,16]. Accessory genes could act as a lineage-specific skill set with an adaptive impact on the bacterium, involved in the colonisation of a novel ecological niche. Genome-wide speciation models, based on ecological or genetic isolation, are mainly focused on core genes[17] but the analysis of lineage-specific genes can provide pivotal information about the emergence of separated genetic clusters within a species.

Serratia marcescens is a Gram-negative opportunistic bacterial pathogen able to cause large outbreaks, in particular in Neonatal Intensive Care Units (NICUs). The bacterium can also be commonly isolated from a multitude of environmental sources, including animal vertebrates[18], insects[19], plants[20], soil[21], and aquatic environments[22,23]. Several evidence of plant-growth promoting activity[20,24] further emphasise the versatile lifestyle of this bacterial species. Despite the health concern, only a few large genomic studies about *S. marcescens* are present in literature and the evolution of this species has been poorly investigated so far. During the last years, the first genomic studies about the *S. marcescens* population structure[25–28] revealed the existence of a certain number of well-defined clades. The most recent and comprehensive studies[27,28], focused mainly on the distribution of clinically relevant features, proposed the existence of one or more specific hospital-adapted lineages, harbouring antibiotic resistance and/or virulence markers. Moreover, a recent wide genomic study on the whole *Serratia* genus[29] has highlighted numerous events of niche specialisation associated with specific gene composition, suggesting that the strong ecological plasticity in the genus is fostered by events of gene gain and loss. Although these studies are progressively shedding light on the population structure and main genomic features of *S. marcescens*, many facets about which mechanisms have played a role in the origin and maintenance of this genetic diversity are still unclear.

The aim of this study was to characterise the diversity within *S. marcescens* and to trace signals of how ecology and gene flow affect the population structure of this wide-spread, ubiquitary, and versatile bacterial species.

Results
Reconstruction of the study Global genomic dataset
The forces shaping the evolution of *Serratia marcescens* were investigated on a large and cured high-quality genomic dataset (labelled Global dataset) including 902 genome assemblies. The genomes were selected from a preliminary collection of 1113 genomes (see "Methods" section, Supplementary Data 1 for details)). The Global dataset comprises: (i) 230 *S. marcescens* genomes sequenced as part of a large study involving six hospitals in Northern Italy[30]; (ii) five additional strains from the same collection of isolates (sequenced ex novo); (iii) 667 genomes from public databases. Overall, this is one of the widest genomic datasets analysed in a comparative genomic study on *S. marcescens* so far. Genomes were manually classified into three categories on the basis of the isolation source: 715 clinical, 122 environmental and 29 animal. For 36 strains it was not possible to obtain a

reliable classification due to the incompleteness of the related metadata. It must be noted that, as in most studies involving opportunistic pathogens, strains from clinical samples are overrepresented in the dataset.

The *Serratia marcescens* population structure reveals five phylogenetic clusters
In the first step of the study, the population structure of *S. marcescens* was investigated by combining core Single Nucleotide Polymorphisms (SNP)-based phylogenetic analysis with Principal Coordinates Analysis (PCoA) clustering on coreSNPs and Mash distances. The SNP calling procedure returned a total of 22,290 coreSNPs and the relative rooted Maximum Likelihood (ML) tree is shown in Fig. 1a. The unsupervised K-means clustering performed on patristic distances, coreSNPs or Mash distances converged in dividing the *S. marcescens* population in five well-distinguished clusters (Fig. S1). The clusters are coherent with the phylogenetic clades (Figs. 1a and S2) and demarcated by deep divisions in the tree. Despite Cluster 1 comprises 53% (475/902) of the strains within the Global genomic dataset, the distribution of Average Nucleotide Identity (ANI) between strains of the same cluster[31], shows that Cluster 4 and Cluster 5 contain clearly more genetic variability than the other clusters (Fig. 1b). Interestingly, ANI among strains of different clusters draws near (and in some cases exceeds) the 95% ANI-based species boundary[32]. Indeed, the maximum ANI between the clusters ranges from 96.78% for the Cluster 1–Cluster 4 pair to 95.56% for Cluster 3–Cluster 4 (Fig. 1c). Overall, the population structure of *S. marcescens* reveals the existence of well-differentiated genetic clusters with clear genetic boundaries, suggesting a remarkable intraspecies genetic diversity.

Specific genomic features highlight diversity between clusters
Genomic features, such as genome size and GC content, were compared between the five clusters (Figure S2). Genome size ranges from 4,955,525 bp to 5,896,859 bp and Cluster 4 has a wider genome size in comparison to Cluster 1, Cluster 2, Cluster 3, and Cluster 5. Cluster 1 genomes are also significantly larger than genomes in Cluster 3. Despite all *S. marcescens* strains displaying a percentage of GC content between 58.9% and 60.2%, comparison between clusters showed that Cluster 1 has a markedly higher GC content than Cluster 2, Cluster 3, Cluster 4, and Cluster 5. At the same time, Cluster 2 also has a lower GC content than Cluster 3, Cluster 4, and Cluster 5. *P* values of significant combinations are shown in Fig. S3 and in Supplementary Note 1.

The synteny analysis performed on 65 complete genomes highlights occasional translocations and inversions occurring among strains of the same Cluster, but synteny is overall preserved in the global population and all clusters share highly syntenic blocks (Fig. S4).

Overall, the observed inter-cluster variations in genome size and GC content are coherent with the cluster's genetic separation described above.

The *Serratia marcescens* clusters are enriched in specific isolation sources
As expected for a human-associated wide-spread bacterium, all clusters are dispersed in every continent apart from Africa and Oceania, greatly underrepresented in the dataset. However, χ^2 test has revealed an uneven distribution of the clusters in the main continents ($\chi^2 = 87.776$, df = 12, p-value = 1.329e-13) and the analysis of the residuals showed that Cluster 5 is associated with North America and negatively associated with Europe (Fig. S5). Moreover, a focus on the spatio-temporal distribution of the 235 strains sampled from six Italian hospitals showed that multiple clusters often coexist within the same hospital in the same time period (Fig. S6).

The geographically balanced analysis of the association between cluster and isolation source (see "Methods" section) indicates that Cluster 1 is significantly associated with clinical settings and negatively associated with environmental sources (Fig. 2). Despite not reaching statistical significance, Cluster 3 and Cluster 5 also display a clear pattern of enrichment in environmental (Cluster 3 and Cluster 5) and animal sources (Cluster 5).

Fig. 1 | The population structure of *Serratia marcescens*. a SNP-based Maximum Likelihood (ML) phylogenetic tree of the 902 *Serratia marcescens* strains of the Global genomic dataset. The tree branches' colours indicate the five clusters coherently and independently determined applying K-means clustering on patristic distances, coreSNP distances and Mash distances. The circle around the tree indicates the strain isolation source (blank if not traceable from the metadata). Bootstrap values are shown on the tree nodes. **b** Distribution of Average Nucleotide Identity (ANI) between *S. marcescens* strains *within* each cluster. The dark blue vertical lines indicate the species identity threshold (95% ANI). **c** Distribution of Average Nucleotide Identity (ANI) *within* and *between* clusters. The dark blue vertical lines indicate the species identity threshold (95% ANI).
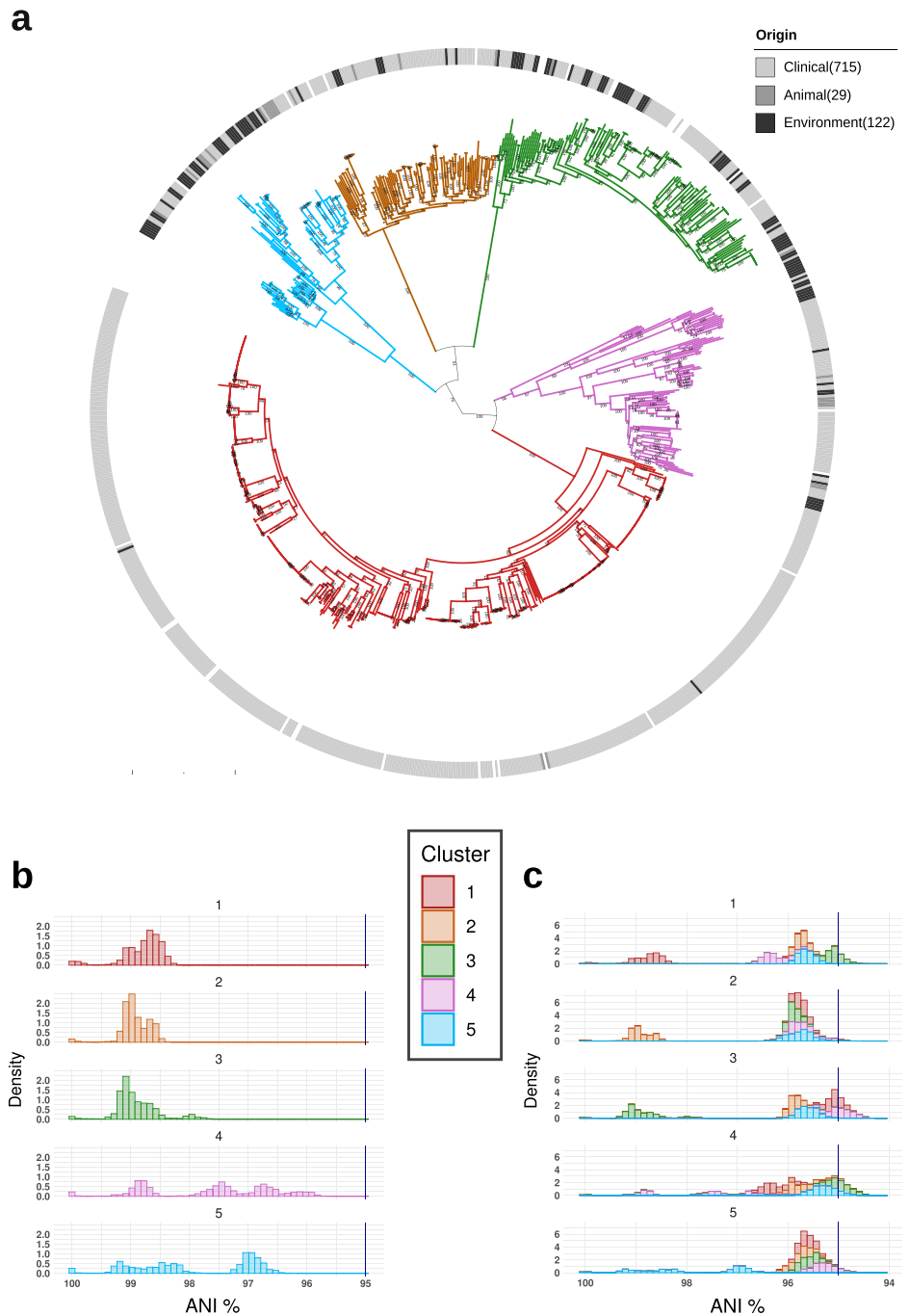
In 2018, Abreo and Altier[25] proposed that *S. marcescens* could be differentiated in an environmental clade and a clinical clade. Following studies on larger genomic datasets have refined this idea, suggesting the existence of one or more clinical/hospital-based lineages[27,28]. Our results highlight that certain clusters are enriched in environmental, animal, or clinical samples, thus providing a signal of possible ecological specificity of the *S. marcescens* clusters. At the same time, different clusters were frequently isolated from the same hospital in the same period, strongly suggesting that the observed genetic separation cannot be explained only by habitat segregation.

Two *Serratia marcescens* clusters have unique gene repertoires

The *S. marcescens* pan-genome comprises a total of 57,700 genes: 2811 core genes (present in ≥95% strains), 3286 shell genes (≥15% and <95%), and 51,603 cloud genes (<15% of the strains). The pan-genomes of *S. marcescens*

and of each single cluster are open (slope of the log-log cumulative curve linear regression < 1, *p* value < 0.05, Figs. S7 and S8). The five clusters show pan-genomes of different size: Cluster 4 exhibits the largest pan-genome and boosts the species total pan-genome, followed by Cluster 5, Cluster 2, Cluster 3, and lastly Cluster 1. Often, the size of a bacterial pan-genome is considered to be related to the lifestyle of the species, and open pan-genomes are associated with ubiquitary bacterial species with wide ecological niches and high rates of horizontal gene transfer[33]. As shown in Fig. 3a, the intensity of gene gain/loss mapped on the phylogenetic tree shows that Cluster 1, Cluster 2, and Cluster 3 exhibit an extensive gene gain/loss on their basal node. Interestingly, major gene gain/loss is also frequent within smaller lineages, reinforcing the assumption that *S. marcescens* undergoes frequent gene turnover.

PCoA on gene presence absence (Fig. 3b) clearly groups the strains coherently with the phylogenetic clusters and the K-means unsupervised

Fig. 2 | Association between phylogenetic cluster and isolation source. Association of *S. marcescens* clusters with animal, clinical or environmental isolation sources. To avoid potential biases due to sampling proximity, χ^2 test was repeated 1000 times on geographically-balanced subsets. The box plot illustrates the distribution of the Pearson residuals for each cluster- isolation source combination, and the red horizontal lines demarcate the thresholds for statistical significance of the residual (see "Methods" section): Cluster 1 is associated with clinical samples (>95% of subsets are significant). Cluster 3 is enriched in environmental isolation sources (52% of subsets), and Cluster 5 is enriched in both environmental (76% of subsets) and animal (42% of subsets). The number of strains in each cluster for each isolation source (*n*) is displayed on the *x*-axis of the boxplots.
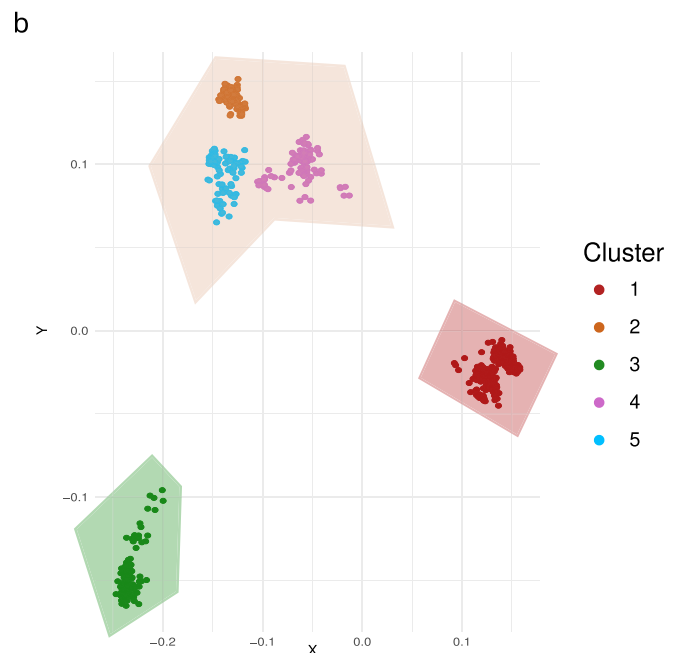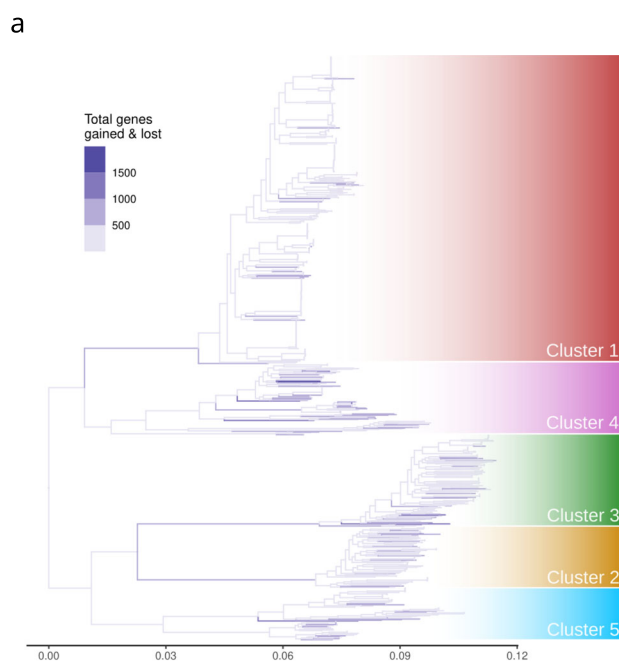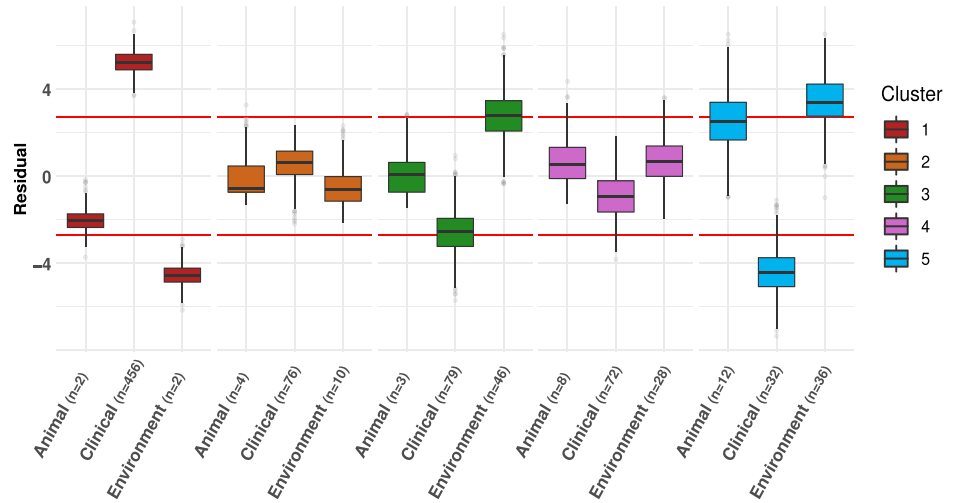

Fig. 3 | Analysis of the *Serratia marcescens* gene repertoire. a Phylogenetic tree with branches coloured according to the number of gene gains and losses inferred by Panstripe. **b** Principal Coordinate Analysis (PCoA) of *S. marcescens* strains based on gene presence absence. Each dot corresponds to a *S. marcescens* strain coloured on the basis of its cluster. The shaded regions represent the three clusters identified by the K-means unsupervised algorithm.

clustering separates Cluster 1 and Cluster 3, grouping together Clusters 2, 4, and 5 (Figs. S2 and S9). This result shows that the five *S. marcescens* clusters have distinct gene content, and Cluster 1 and Cluster 3, previously found to be associated with clinical and environmental sources, are remarkably different from the others.

A more in-depth analysis identified 107 genes specific for Cluster 1 (i.e. present in >95% of Cluster 1 strains and <15% of the strains of the other clusters), 58 genes for Cluster 2, 168 for Cluster 3, 14 for Cluster 4, and 81 for Cluster 5. COG-annotated Cluster-specific genes are available in Supplementary Data 2.

In summary, at the beginning of their separation, three clusters underwent frequent episodes of gene gain/loss and two of these clusters (Cluster 1 and Cluster 3) reached a unique gene repertoire. Since these clusters were notably found to be enriched in clinical and environmental samples, their gene repertoire is coherent with independent adaptive trajectories towards specific lifestyles. Despite being grouped with Cluster 4 and Cluster 5, also Cluster 2 displays a clear pattern of differentiation in gene content.

The habitat of *Serratia marcescens* clusters inferred from shotgun metagenomics analysis

As stated above, some clusters present a clear enrichment for specific isolation sources, such as the Cluster 1 for the hospital settings and Cluster 3 for the environment. However, *S. marcescens* is mainly studied for its clinical relevance, producing a strong sampling bias towards hospitals and human samples. To overcome this limit, we investigated the presence of strains of the *S. marcescens* clusters in different biomes using a large metagenomics database.

Firstly, we identified protein markers specific to *S. marcescens* and others able to distinguish the clusters. As to *S. marcescens* protein markers, the 40 *S. marcescens*-specific proteins found by Alvaro and colleagues[30] were tested and 27 resulted to be discriminant. To distinguish the *S. marcescens* clusters, the cluster-specific proteins found above were tested: 46 gene markers were selected for Cluster 1, 11 for Cluster 2, and 20 for Cluster 5. For Cluster 4 and Cluster 3 it was not possible to identify reliable markers. For Cluster 4, the lack of protein markers can be explained by the fact that only

Fig. 4 | Inference of *S. marcescens* clusters in metagenomic samples. Cluster-specific core genes with high rates of specificity and sensibility (see "Methods" section) were searched on metagenomic samples of the MGnify database to identify phylogenetic clusters of *S. marcescens* in a broad set of biomes. A metagenomic assembly was considered positive to a Cluster if >10% of cluster-specific core genes were found within it. When a metagenomic assembly resulted positive to core *S. marcescens* genes, but not to genes associated to Cluster 1, Cluster 2 or Cluster 5, the assembly was defined as Other SMA. The heatmap shows the residuals of the χ^2 test used to investigate whether *S. marcescens* clusters are associated to samples from specific biomes. Statistically significant associations are marked with asterisks ('∗', *p* value < 0.05). Here, only biomes with more than 100 positive samples are shown, see Fig. S11 for all biomes.
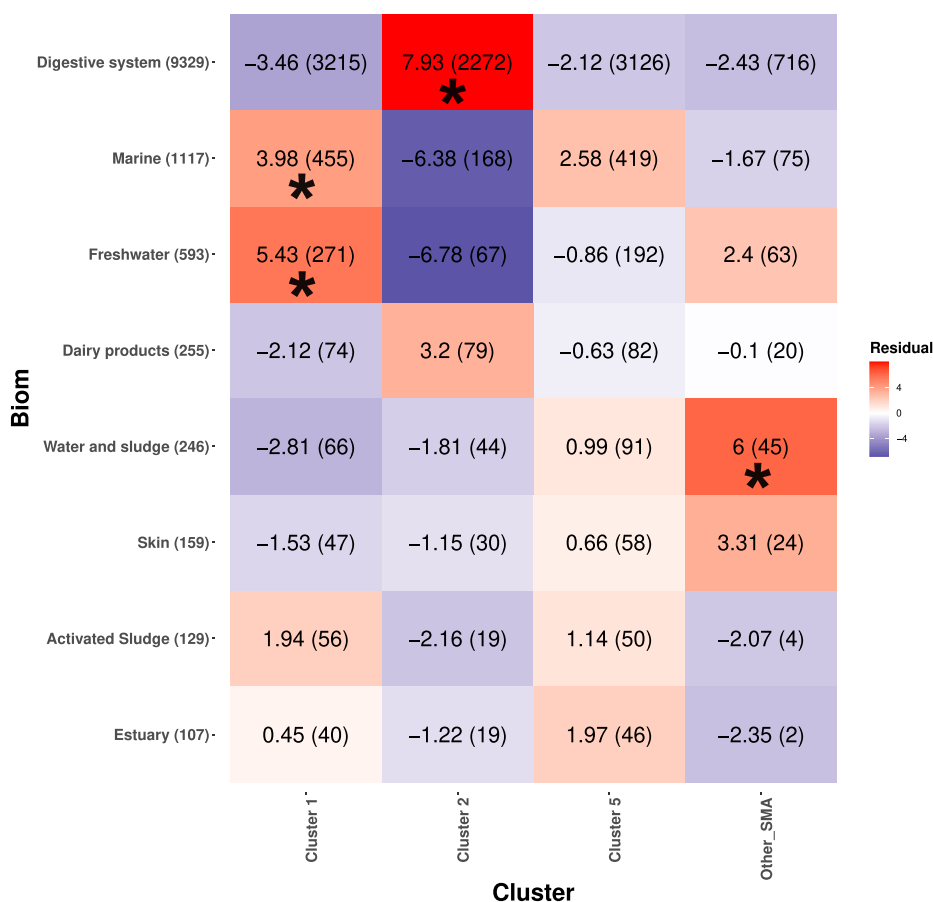
14 cluster-specific core genes were identified (see the Specific gene repertoires suggest clusters ecological adaptations section). On the other hand, the absence of specific genes for Cluster 3 can be explained considering the high similarity of its cluster-specific genes with those of other bacterial species (even outside the *Serratia* genus, see Figure S10). This suggests that the separate gene content of Cluster 3 could arise from intense gene flow with other bacterial species, coherently with the recently proposed idea that the evolution of the *Serratia* genus is affected by interspecies gene flow[29].

To study the distribution of the *S. marcescens* clusters the protein markers were searched into MGnify[34], a large database containing hundreds of thousands of protein sequences from shotgun metagenomics data on several biomes. The search of *S. marcescens*-specific protein sequences into the MGnify database identified a total of 6235 metagenomic-based assemblies possibly containing *Serratia marcescens* sequences. Among these *S. marcescens*-positive assemblies, 5250 (84%) resulted positive to at least one *S. marcescens* cluster, and 1604 (31%) resulted positive to a single cluster. Despite a general biome co-presence was observed for the clusters (Fig. S11), some interesting statistically significant associations were observed: Fig. 4, which only takes most relevant biomes into account, shows that Cluster 1 was enriched in aquatic biomes (i.e. marine and freshwater) and Cluster 2 in the digestive system. (For details on all biomes where *S. marcescens* was traced see Fig. S12).

This analytical approach presents some issues. Despite the target proteins being selected on the basis of their high specificity, HGT events among *S. marcescens* clusters and between *S. marcescens* and other species cannot be excluded. Indeed, the used protein markers have a sensibility/specificity threshold of 75%. Moreover, metagenomic datasets are highly susceptible to chimeric sequences assembly[35]. Nevertheless, this analysis represents a useful tool to broaden our knowledge on the habitat colonised by *S. marcescens*. It underlines the ecological plasticity in *S. marcescens* and fortifies the idea that, although wide-spread and often co-existent, clusters could have individual ecological preferences. It is of particular interest that Cluster 1 (strongly associated with clinical samples and harbouring virulence factors) was found to be enriched in freshwater, suggesting a possible reservoir for this pathogenic bacterium.

Reconstruction of the Refined genomic dataset and phylogenetic tree

The analyses used to investigate the *S. marcescens* clusters origin and maintenance (recombination analysis, molecular clock, and HGT analysis) are sensitive to genetic dataset biases. For this reason, the Global genomic dataset was refined to balance the *S. marcescens* genetic variability (see "Methods" section and Fig. S13). The obtained Refined genomic dataset included a total of 86 representative strains: 19 from Cluster 1, 16 from Cluster 2, 12 from Cluster 3, 21 from Cluster 4, and 18 from Cluster 5. Then, the ML phylogenetic tree was built on the relative 365,317 coreSNPs. The obtained tree was globally coherent with that obtained on the Global genomic dataset and all the clusters corresponded to monophyletic highly-supported groups (bootstrap supports 100, Fig. S14).

Large genomic recombinations contributed to *Serratia marcescens* diversification

The Refined dataset whole-genome alignment and the relative ML phylogenetic tree were subjected to recombination analysis to investigate its impact on the evolution of *S. marcescens*. As a whole, *S. marcescens* exhibited a recombination to mutation ratio (r/m) ratio of 2.35, being significantly less recombinogenic than what estimated for opportunistic pathogens[36] such as *Salmonella enterica* (r/m = 30.2), *Streptococcus pyogenes* (r/m = 17.2) and *Helicobacter pylori* (r/m = 13.6). Still, this r/m value is comparable to other opportunistic pathogens like *Campylobacter jejuni* and *Haemophilus parasuis*, suggesting that homologous recombination is implicated in the shaping of genetic diversity within *Serratia marcescens*. Furthermore, large
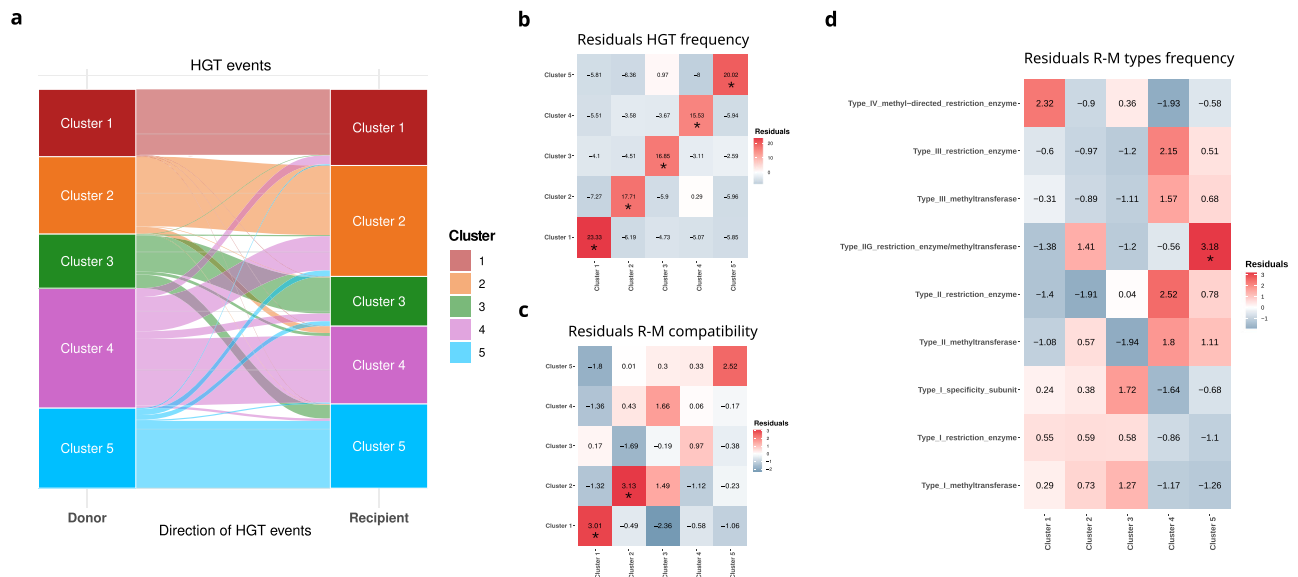
Fig. 5 | Gene flow and Restriction-Modification (R-M) systems compatibility.
a Stacked bar charts showing the percentage of genes with Horizontal Gene Transfer (HGT) events inferred for each cluster from 1062 core genes. Intra indicates genes with HGT events which happened within strains of the same cluster. Donor indicates genes with HGT events departing from the cluster to other clusters. Recipient indicates genes with HGT events departing from other clusters. **b** Heatmap showing the χ^2 residuals of *S. marcescens* clusters pairs on the basis of the total number of HGT events inferred between them. Statistically significant associations are marked with an asterisk (*). **c** Heatmap showing the χ^2 residuals of *S. marcescens* clusters pairs on the basis of R-M systems compatibility. Statistically significant associations are marked with an asterisk (*). **d** Heatmap showing the χ^2 residuals of R-M system type distribution in the S. marcescens. Statistically significant associations are marked with an asterisk (*).

recombinations (> 100,000 pb) were mapped on basal nodes of Cluster 2 and Cluster 3, suggesting that the divergence among these two clusters emerged in correspondence of major recombination events. Large recombinations are also evident within Cluster 4 and Cluster 1. Recombination parameters were estimated for each branch of the tree and the distribution of the r/m ratio within the five clusters were compared: Cluster 2 has the highest distribution of r/m ratio and is significantly more recombinant than Cluster 3 (*p* value < 6.4e-07), Cluster 4 (*p* value < 4.3e-09), and Cluster 5 (*p* value < 0.00015). Cluster 1 and Cluster 5 are also significantly more recombinant than Cluster 4 (*p* value < 0.00693 and *p* value < 0.01138). Large recombinations along the phylogenetic tree and r/m ratios for each cluster are shown in Fig. S15.

Interestingly, the recombinations were not randomly scattered along the genome but there is a 10,000 bp long hyper-recombinated region. This region contains the capsular genes *wza*, *wzb* and *wzc* of the *wz* operon, and a phylogenetic reconstruction of their concatenate has confirmed that they are highly recombined (Fig. S16). The bacterial capsule is a well-known virulence factor[37,38] and capsular locus have been shown to be recombination hotspots as consequence of immune escaping[39,40]. This suggests a dynamic interaction with other organisms for all clusters, but could also be linked to the ability to colonise and adapt to diverse ecological niches[41].

Clusters exhibit limited genetic exchange and preferential gene flow routes

Up to here, it was established that the *S. marcescens* population is divided in five well-delimited clusters, emerged also by large recombinations and having specific genetic features, including gene content and recombination rate. To unveil whether preferences in genetic exchange could be involved in the maintenance of genotypic clusters within the species, gene flow within *S. marcescens* was investigated.

The HGT analysis performed on the 1062 core genes identified 676 events on 443/1062 (42%) genes. More in detail, 517/676 (76%) HGT events, occurred on a total of 374/443 (84%) genes, involved strains of the same cluster, while 159/676 (24%) HGT events, occurred on a total of 142/443 (32%) genes, involved strains of different clusters (Fig. 5a). Among the 517 intra-cluster HGT events, 111/517 (21%) were within Cluster 1, 118/517

(23%) within Cluster 2, 61/517 (12%) within Cluster 3, 113/517 (22%) Cluster 4, and 114/517 (22%) within Cluster 5. The inter-cluster HGT events involved preferentially specific pairs of clusters: most cluster pairs exchanged maximum 1% of the 1062 core genes, while Cluster 2–Cluster 4 pair exchanged 68 core genes (>6%) and Cluster 3–Cluster 5 pair exchanged 31 genes (>3%). The preferential trend towards intra-cluster HGT is also evident from the residuals of the χ^2 test (Fig. 5b).

Interestingly, Cluster 1 (associated with clinical samples and a unique gene repertoire) has the most limited genetic exchange with the other clusters. On the other side, Cluster 4 shows a notable genetic exchange with Cluster 2, despite the phylogenetic distance. Altogether, these analyses revealed a limited gene flow among the clusters, suggesting an absence of genetic exchange coherent with the genetic separation described above.

Restriction-Modification (R-M) systems could be involved in the genetic isolation of *Serratia marcescens* clusters

One of the main gene flow barriers is the incompatibility between Restriction-Modification (R-M) systems. Bacteria modulate the acquisition of foreign DNA (i.e. avoiding phagic DNA) using two-components Restriction-Modification (R-M) systems: the first enzyme (a methylase) methylates specific DNA patterns, while the second enzyme (an endonuclease) cleaves DNA when the same pattern on the DNA is not methylated. Thus, two bacteria can exchange DNA more successfully if they harbour compatible R-M systems. Genetically close strains tend to harbour similar R-M systems, thus similar strains were grouped and only one representative strain per group was included in the R-M graph (see Methods). A total of 84 strains were selected for R-M graph reconstruction: 30 from Cluster 1, 10 from Cluster 2, 16 from Cluster 3, 20 from Cluster 4, and 8 from Cluster 5. The R-M compatibility among the *S. marcescens* strains of the Global dataset was studied and the R-M compatibility among the clusters was studied by χ^2 test. The analysis revealed that Cluster 1, Cluster 2 and Cluster 5 have a strong intra-cluster preferential R-M compatibility, as shown by the residuals in Fig. 5c. As shown in Fig. 5d, χ^2 test shows that there are R-M systems enriched in *S. marcescens* clusters (see also Supplementary Data 3). Notably, Cluster 5 is associated to Type IIG R-M systems and clinical-associated Cluster 1 is exclusively enriched in Type IV R-M systems. Furthermore, χ^2

Test residuals obtained by the analysis of R-M systems resulted significantly correlated to those obtained by HGT analysis (Linear regression, $p = 0.0008$). This result suggests that the observed genetic isolation of the clusters could be partially explained by genetic barriers due to R-M system incompatibility.

Discussion

Serratia marcescens is an infamous nosocomial pathogen able to cause large and fatal outbreaks in Neonatal Intensive Care Units (NICUs) and to rapidly spread in hospital settings[42,43]. The bacterium is also able to colonise soil, water, plants, and animals such as insects and corals. Recent genomic studies have identified clades associated with different isolation sources, including clinical/hospital-based lineages harbouring several virulence/resistance traits[25–29].

In this work, we investigated the diversity of this bacterium, with a strong focus on the genetic/ecological mechanisms underlying its population structure. Our results clearly showed that *Serratia marcescens* is composed of five well-defined major genetic clusters, exhibiting specific gene content and limited inter-cluster gene flow. The phylogenetic tree topology is mainly coherent with two recent works on *S. marcescens*[27,29] and PCoA-based unsupervised K-means clustering led to the identification of major large-scale clusters. However, as shown in literature, the population structure could be divided into more sub-clusters[28,44], varying in number on the basis of dataset and clustering algorithm. Considering that the aim of this work was to study the major evolutionary forces shaping *S. marcescens* population structure, we decided to focus on the five major clusters.

At least two clusters also showed genetic signatures of ecological adaptations. Cluster 1 is frequently associated with hospital settings: it has one of the most reduced pan-genome sizes, high rates of gene gain/loss in its phylogenetic basal node and it comprises a very specific gene repertoire, including genes involved in virulence mechanisms. Among the genes associated to Cluster 1 (see Supplementary Data 2) there are *fhaC*, with a role in the haemolytic process, *fimC* and *hrpA*, both involved in fimbrial biogenesis and *lacZ*, linked to coexistence with mammals[45]. Interestingly, the presence of *fhaC* is essential for pathogenicity in *Bordetella pertussis*[46]. This cluster also has reduced intra-cluster genomic variability, a very limited gene flow with other clusters. Overall, these genetic features are compatible with an association to a hospital-related lifestyle. Interestingly, metagenomics analysis revealed that Cluster 1 could use water (marine or freshwater) as its reservoir habitat. This environmental association suggests a possible explanation of the ability of strains belonging to this cluster to rapidly spread in the hospital environment and to colonise several substrates, but further studies are required to test this hypothesis. Cluster 2 also displays several events of gene gain and loss at its root, which produced a gene composition distinct but not so distant from Cluster 5 and Cluster 4. It is not associated with any isolation source, but the metagenomic analysis suggests its enhancement in the human digestive system. Cluster 3 is enriched in environmental sources (such as soil, plants, and water) and displays a specific gene repertoire which includes genes involved in the metabolism of plant and fungal carbohydrates (*pulE*, *pulC*, *pulK*, and *pulL*). These four genes are involved in the metabolism of plant and fungal carbohydrates, and perhaps could provide a signal of adaptation to plant environments[47]. The cluster has also a reduced genetic variability and it displays signals of an extensive gene flow with other bacterial species. Cluster 4 is not associated with any isolation source. It is characterised by large genetic variability, greater genome size, and the widest pan-genome, with very few specific genes. Cluster 5 is enriched in environmental and animal sources but the genetic variability within the cluster is high and gene composition is very similar to Cluster 4. Moreover, the metagenomic analysis did not reveal a specific association for any biome. Interestingly, it is the only cluster with a macro geographic uneven distribution (association to North America).

The existence of deeply demarcated clusters in *S. marcescens* suggests geographical, ecological, or genetic barriers behind the origin and maintenance of this diversity. The analysis of isolation sites strongly supports the absence of geographical barriers: strains of all the clusters have been often isolated in the same hospital in the same period of time, in our dataset as well as in the study published by Moradigaravand and colleagues[44].

Considering this limited geographical isolation, the potential ecological divergence between the clusters was investigated using genomic and habitat information. More in detail, gene content of each cluster was compared to trace signals of ecological adaptation[8] and, when available, specific genetic traits were searched in a large metagenomic database to better understand cluster habitat and evade sampling biases. Interestingly, Cluster 1 and Cluster 3 showed consistent signals coherent with independent adaptive trajectories to hospital and environmental settings, respectively. This result partially recalls the idea first proposed by Abreo and Altier in 2019[25], that *S. marcescens* has diverged into an environmental clade and a clinical clade. However, unlike suggested by Abreo and Altier, the divergent ecotypes represent two emerging clades and only represent a minimal portion of the genetic variability observed in *S. marcescens*.

The absence of strongly evident ecological adaptation for each of the clusters led us to investigate the origins of these clusters in light of recombinations and gene flow. The clusters mainly show an average recombination rate, in comparison to other species, even if several large recombination events were detected, within the clusters but also on the basal nodes of sister-groups Cluster 2 and 3. The reconstruction of horizontal gene transfer (HGT) events on core genes revealed a very limited inter-cluster gene flow, suggesting the presence of ancient and strong barriers to recombination. Furthermore, the analysis of Restriction-Modification (R-M) systems revealed a partial incompatibility among clusters. It was also noted that clusters are enriched in different types of R-M systems and it is widely recognised that similarity and compatibility of R-M systems between strains promote HGT and vice versa[48]. Indeed, studies show that type I and type III-like R-M systems can act as genetic barriers in *Paenibacillus polymyxa*[49], *Enterococcus faecium*[50], and *Staphylococcus aureus*[51,52]. We found that Type IV R-M systems are exclusively enriched in Cluster 1. This R-M system type is known to cleave modified DNA sequences, limiting the acquisition of foreign DNA[53]. Thus, the presence of this R-M system could contribute to the genetic isolation of Cluster 1. It also is important to note that R-M systems are not the only barrier to horizontal gene transfer and other factors such as DNA sequence similarity, limits on host transfer range, maintenance mechanism of mobile genetic elements are involved[10,11]. Indeed, Cluster 3 and Cluster 4 still exhibit preferential intra-cluster gene flow despite a low intracluster R-M compatibility. Taken together, these results suggest that genetic barriers and a lack of genetic exchange have had a major role in the divergence of the clusters, shaping the current population structure of *S. marcescens*.

In conclusion, *S. marcescens* is composed of five major clusters separated by strong genetic barriers. Within this population structure, two clusters (Cluster 1 and 3) have initiated adaptive trajectories to specific ecological niches and proceed to progressively isolate from the others. Whereas, other clusters are ecologically generalist and despite they often co-occur in the same environment at the same time, genetic barriers are sufficiently thick to maintain the clusters regardless of ecology or spatial distribution. Thus, with a hint of speculation, we propose that the leading role in the evolution of *S. marcescens* is played by the genetic barriers between co-occurring, ecologically generalist subpopulations. Eventually, due to environmental pressure and constant reshuffling of the accessory genome with other species, adaptive populations have emerged. Our results open to interesting biological questions, such as: what caused the arisal of genetic barriers in the first place? Is this cluster-like population structure in equilibrium or are the adaptive clusters embarked on paths towards speciation? At what point could these clusters be considered as subspecies?

Methods
Global genomic dataset preparation
The preliminary genomic dataset used in the study contained a total of 1113 genome assemblies: (i) 871 *Serratia marcescens* genomes available on March 10, 2022 in the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) for which geographical information and isolation date were reported;

(ii) seven additional genomes used in previous genomic studies[25,26,44] and absent in the BV-BRC database; (iii) 230 *S. marcescens* genome assemblies sequenced in the previous study[30]; (iv) five *S. marcescens* genome assemblies of strains isolated from the Italian hospitals ASST Papa Giovanni XXIII Hospital in Bergamo ($n = 2$), RCCS San Raffaele Hospital (HSR) in Milan ($n = 2$) and ASST Fatebenefratelli Sacco Hospital in Milan ($n = 1$) (details about genome sequencing and assembly are reported below). Details are reported in Supplementary Data 1.

Five *S. marcescens* isolates were grown on McConkey agar medium overnight at 37°C. The day after, single colonies were picked and DNA extractions were carried out using a Qiagen QIAcube Connect automated extractor (Qiagen, Hilden, Germany) following the bacterial pellet protocol which employs Qiagen DNeasy Blood & Tissue reagents. Then, libraries were prepared and 2×150 bp paired-end run sequencing was carried out on the Illumina NextSeq platform. The reads were quality checked by using FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then assembled using SPAdes[54].

Within the preliminary genome dataset, the low quality genome assemblies and those for which the *S. marcescens* taxonomy was incorrect were detected and removed to obtain the Global genomic dataset.

The assembly quality parameters used for the selection were: assembly total length, number of contigs, N50, N count and the Open Reading Frame (ORF) number. ORF calling was performed using Prodigal[55] and the genome statistics were obtained using the assembly-stats tool (https://github.com/sanger-pathogens/assembly-stats). For each of these parameters, the thresholds for the selection were computed on the starting genomic dataset using the Tukey's fences statistical method[56]: the lower boundary (L) is computed as $Q1 - (1.5 \cdot IQR)$ and the higher boundary (H) as $Q3 + (1.5 \cdot IQR)$, where Q1 indicates the first quantile of the value distribution, Q3 indicates the third quartile and IQR indicates the interquartile range. The obtained thresholds were: (i) Total length between 4,500,000 bp and 6,000,000 bp; (ii) Number of contigs < 116; (iii) N count < 5842 ; (iv) N50 > 7077 ; (v) 5134 < ORF count < 4594. The N count parameter was considered crucial for high-quality and all genomes that did not respect its threshold were excluded. Among the remaining genome assemblies, those that passed at least three out of the other four quality checks were selected for the taxonomy-based step of selection.

Taxonomy of the genomes were assessed combining Average Nucleotide Identity (ANI) and 16S rRNA sequence. The Mash pairwise distance matrix was computed between all genomes using Mash[31] and the genomes were clustered with a cut-off distance of 0.05. The 16S rRNA sequence was extracted using Barrnap and Blastn-searched into the 16S rRNA database Silva[57]: the genomes were then classified on the basis of the best hit as *Serratia marcescens*, *Serratia* spp. and Others. The 16S rRNA gene is in multiple copies within the *S. marcescens* genome making it difficult to assemble. Genomes for which it was not possible to identify the 16S rRNA gene were classified as undefined. Combining the Mash-based clustering and the 16S rRNA classification, a *Serratia marcescens*-like cluster was defined. The genomes clustered within the *Serratia marcescens*-like cluster and annotated with 16S rRNA as *Serratia marcescens*, *Serratia* spp. or undefined were selected. Herein, the selected genome dataset will be referred to as the Global genomic dataset.

Genome classification by origin
Based on the sampling material, *S. marcescens* genomes were manually distinguished into three ecological categories: (i) clinical, if the sample was obtained from a clinically-related human sample; (ii) animal, if the bacterium was associated to any non-human metazoan; (iii) environmental, if the sample was found on any other environmental source, such as water, plants, and soil.

Population structure
The assemblies of the Global genomic dataset and one outgroup (*Serratia plymuthica* strain 4Rx13, GCF_000176835.2) were aligned against the genome assembly of the reference strain *S. marcescens* Db11 and Core Single

Nucleotide Polymorphisms (CoreSNPs) were called using the tool Purple[58]. The obtained CoreSNPs were subjected to Maximum Likelihood (ML) phylogenetic analysis using FastTree MP[59] (with 100 pseudo-bootstraps), using the general time reversible (GTR) model. The obtained tree was manually rooted on the outgroup using Seaview[60]. Lastly, the web-based tool iTOL[61] was used to map strains metadata on the topology.

The global genomic dataset *strains were* grouped via Principal Coordinates Analysis (PCoA) and unsupervised clustering algorithm K-means, using independently tree patristic distances, CoreSNP distances, Mash distances and Jaccard distances computed on the gene presence absence. The Average Nucleotide Identity (ANI) between strains was computed as $(1 - \text{Mash distances}) \times 100$[31]. For each analysis, the optimal number of clusters was determined in accordance to the best average silhouette score.

Clusters comparison
Genome size, number of genes and GC content were compared between clusters by Mann–Whitney *U*-test with Holm post-hoc correction and visualised by boxplots. Pairwise SNP-distances and ANI distances were used to infer genetic diversity among strains and compared among the clusters by histograms. The analyses were performed using R.

The genomic synteny within and between clusters was assessed on the 65 complete genomes available within the global genomic dataset. Before the analysis, plasmidic contigs were manually removed and the chromosomes were re-arranged on the basis of the *dnaA* gene position. For each cluster the re-oriented genome assemblies were aligned using progressiveMauve[62] and the intra-cluster synteny plot was obtained using the R package genoplotR[63]. The inter-cluster synteny was investigated using one representative strain per cluster.

The geographic distribution of the strains of the different clusters was compared using the χ^2 Test of Independence on the isolation continents. Pearson's standard residuals were evaluated to investigate geographic distribution of the clusters: i.e. residuals were considered as statistically significant when the value was greater than the Bonferroni-corrected critical value[64].

As stated above (i.e. section Genome classification by origin), the *S. marcescens* strains of the Global genomic dataset were assigned to ecological categories on the basis of their isolation source. To investigate ecological preferences among clusters, the Pearson's standard residuals of the χ^2 Test of Independence between *S. marcescens* cluster and the relative ecological categories were studied. The residuals were considered as significant if their absolute value was greater than the Bonferroni-corrected critical value. To minimise the possible bias due to geographical proximity of the samples, a geographically-balanced χ^2 Test of Independence was implemented with a Monte Carlo method: the test was run 1000 times, sampling 40 genomes from each continent. A cluster was considered statistically associated to a specific ecological category when the relative standard residual was significant in at least 950 test runs out of 1000 replicates. Strains from Africa ($n = 8$) and Asia ($n = 3$) were excluded from the analysis because of the very low representation of these continents in the dataset; North America ($n = 96$) and South America ($n = 22$) were merged into the America.

Genomes were annotated using Prokka[65] and General Feature Format (GFF) files were fed to Roary[66] for pan-genome analysis. Pan-genome cumulative curves were built using R for the entire dataset and for each cluster independently. Then the open vs close status of each pan genome was assessed as described by Tettelin et al.[67].

Gene gain and loss events were mapped on the tree with Panstripe[68], using maximum parsimony as method for the ancestral state reconstruction.

Differences in gene content among clusters were also investigated by PCoA on the gene presence/absence Jaccard distance matrix obtained from the Roary tool.

Orthology groups that were found to be core (> 95% present) in one cluster and rare in all other clusters (<15% present) were considered cluster-specific core genes. Nucleotide distances among sequences of each orthology group were computed via the *dist.dna* function of the ape R library[69], and the

sequence with the lowest mean nucleotide distance from the others was selected as representative of each orthology group. Representative sequences were annotated against the COG-database using the tool COGclassifier (https://pypi.org/project/cogclassifier/). Moreover, genes were defined as chromosomal or plasmidic by BLAST search against the complete genomes with plasmids included in the dataset.

The analysis of ecological enrichment performed on the Global genomic dataset (see above) can suffer from sampling bias. Indeed, as expected, most of the strains were isolated from clinical settings. The Mgnify[34] protein database contains protein sequences obtained from shogun metagenomics sequencing of thousands of samples collected from a vast range of biomes/ecological sources. To assess the presence of sequences specific to the different *S. marcescens* clusters into the samples this database can help to overcome this issue.

To do so, it was necessary to use protein markers able to discriminate the *S. marcescens* clusters from all the other bacterial species. The protein sequences of the cluster-specific core genes (see section Annotation of cluster-specific core genes) were searched by DIAMOND[70] (E-value < 0.00001, sequence identity ≥ 90% and the ratio between query length and length of the hit ≥ 0.85 ≤ 1.1) against all proteins of the genomes of the Global genomic dataset, in order to assess the ability of these target proteins to identify the clusters by DIAMOND search. For each target protein, the sensibility and specificity was evaluated using the Youden's index and the best threshold for the percentage of sequence identity was determined in a range from 90 to 99. The specificity of the target proteins for *Serratia marcescens* was assessed in a similar way: protein sequences were searched by DIAMOND against the NCBI NR[71] database. After filtering for coverage and e-value as above, the specificity of each target protein for *S. marcescens* was calculated on the basis of the hit sequence taxonomy, using thresholds of sequence identity percentages between 90 and 99. The highest value of specificity was extracted together with the corresponding percentage of sequence identity used as threshold. The protein sequences of 40 core genes with a good specificity for *S. marcescens* determined by Alvaro et al.[30] were also included in this analysis. The genes were selected to be appropriate markers only if, at a certain threshold, their Youden's index value was higher than 0.75 and the specificity to *S. marcescens* higher than 75%.

All sequences of the marker genes were searched by DIAMOND against the MGnify protein database. The results were filtered as above for coverage and e-value, while the percentage of sequence identity threshold used was target-specific. A MGnify sample was considered to contain *S. marcescens* if at least three of the selected Alvaro et al.[30] protein targets were present. These samples were further investigated for the determination of the *S. marcescens* cluster present by DIAMOND searching for the cluster-specific protein markers. To determine if *S. marcescens* clusters were linked to different biomes/ecological sources, χ^2 Test of Independence was performed (standard residuals were considered significant if their absolute value was greater than the Bonferroni-corrected critical value).

Cluster gene flow and recombination analysis

The analyses for the investigation of cluster origin and maintenance (including recombination, gene flow and molecular clock analyses) are sensitive to the size and genetic bias of the genomic dataset. To reduce the size of the Global genomic dataset, maintaining the genetic variability as much as possible, the genomes were grouped on the basis of pairwise coreSNP distance: the strains having coreSNPs distance below a specific threshold fell in the same group and the youngest and oldest (on the basis of the isolation date) strains of the group were retrieved. To define the best threshold to be used, the number of groups over SNPs thresholds ranging between 0 and 1000 SNPs were plotted using R. Herein, this dataset will be referred to as Refined genomic dataset.

The Purple[58] tool was used for the reference-based coreSNP calling. The genome of the Refined genomic dataset was aligned to the *S. marcescens* Db11 reference genome assembly and SNP were called and used to obtain the whole-genomes alignment and to extract the coreSNPs[58]. The extracted coreSNPs were then subjected to Maximum Likelihood (ML) phylogenetic

analysis using RAxML8[72], applying a general time reversible model that incorporates rate variation among sites and proportion of invariant sites (GTR + G + I), according to ModelTest-NG[73].

The ML phylogenetic tree and the whole-genome alignment (obtained using Purple[58]) were fed to ClonalframeML[74] for recombination analysis. Recombination events were estimated per-branch and ambiguous sites on the alignment were ignored in the analysis. From ClonalframeML output, r/m ratio was calculated as $r/m = r/theta * delta * vu$ and compared among clusters via Mann-Whitney U-test with Holm post-hoc correction. Then, for each cluster the cumulative number of recombined bases within windows of 5-kbp along the whole-genome alignment was computed.

The gene annotation of reference genome Db11 was checked to identify genes located on highly recombined regions. The ML phylogeny of the genes of interest within the recombined regions were obtained using RAxML8 with 100 pseudo-bootstraps after best model selection using ModelTest-NG. Moreover, single large recombination events along the genome (> 100 kbp) were mapped on the phylogenetic tree.

Core gene alignments were extracted from the whole-genomes alignment obtained above (see SNP calling, SNP annotation and Maximum Likelihood phylogenetic analysis) on the basis of the positions of the Coding DNA Sequences (CDSs) on the reference Db11 *S. marcescens* strain genome. Each gene alignment was subjected to ML phylogenetic analysis using RAxML8 after best model selection using ModelTest-NG. The topology of each tree was compared to the SNP-based phylogenetic tree using T-REX command-line version[75] to detect HGT: the analysis was repeated on bootstrap trees and only HGT events with a bootstrap support of at least 75 were considered reliable. The HGT analysis returns the nodes of donors and recipients of each detected HGT event. Using this information, the network describing the gene flow between clusters in *Serratia marcescens* was constructed using Gephi[76]. Lastly, the preferential association among clusters for HGT events was evaluated analysing the residuals of the χ^2 Test.

The methylase and endonuclease enzymes of the R-M systems present in the strains of the Global genomic dataset were identified and annotated by Blastn search against the REBASE database[77], selecting the best hits with coverage (hit length / query length) > 0.9 and nucleotide identity > 90%. The hits were then classified as Orphan methyltransferase (methylases without the relative endonuclease, usually involved in gene regulation), Methyltransferase and Endonuclease. When the HGT donor strain harbours methylase enzymes compatible with the endonuclease enzymes of the recipient strain (i.e. the two enzymes methylate/not-cleave the same DNA pattern), the transferred DNA is more likely to be incorporated by the recipient. Strains genetically very similar will tend to harbour similar R-M systems, because they share a closer ancestor and it is reasonable that this could lead to the overestimation of the intra-cluster R-M compatibility. To avoid this bias the strains were previously clustered using the coreSNP alignment obtained above with a threshold of 10 SNPs. Among the strains of the same cluster the one harbouring more R-M system genes was selected as representative. These selected strains were then used to reconstruct the R-M graph where the nodes are the strains, and two nodes are connected if all the R-M systems harboured by the strains are compatible. The preferential association of the clusters on the graph was studied by χ^2 Test of Independence.

To investigate whether the R-M system could affect the observed gene flow pattern (as determined above), the χ^2 Test residuals of the preferential association between clusters computed from the R-M graph and the residuals obtained from HGT analysis were compared by linear regression. A χ^2 Test of Independence and an analysis of Pearson's standard residuals was used to investigate whether specific types of R-M systems are associated with clusters.

Statistics and reproducibility

This study was conducted on a dataset of 902 *S. marcescens* genomes.

All statistical analyses were performed using R software. Boxplots were used to visualise distribution of continuous variables, and the combination

of Kruskal-Wallis test and Mann-Whitney U test were used to test differences in the distribution between groups. Associations between categorical variables were tested by χ^2 test and subsequent analysis of Pearson's standard residuals. Since standardised Pearson residuals are normally distributed with a mean of 0 and standard deviation of 1, the critical value N was calculated on the Bonferroni-corrected p-value and residuals with an absolute value above N were considered statistically significant. Repeated sampling of strains stratified by geographical origin was implemented to ensure that associations between phylogenetic clusters and isolation sources were not influenced by geographical proximity of samples.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
The sequencing data generated by this study has been deposited in the NCBI database under Bioproject number PRJNA957961. Other whole genome sequences used in this study were collected from BV-BRC and NCBI, with the accession number for each sequence provided in Supplementary Data 1. Source data is available on Figshare at 10.6084/m9.figshare.24975591[78]. The phylogenetic tree, SNP-alignments and Mash distance matrix underlying Fig. 1 are found in the three files located in folder "1_Phylogeny_and_-clustering". The cluster-isolation source association table underlying Fig. 2 is found in the file "2_Isolation_source_clusters_association/Chi_squared_Pearson_residuals_1000_repetition.tab". The Roary gene presence absence matrix underlying Fig. 3 is found in "3_Gene_association/gene_-presence_absence.Rtab". The MGnify search output table underlying Fig. 4 is found in "4_Search_cluster_specific_genes_on_MGnify/MGnify_-Bioms_Sma.tab". The table with HGT events and the table of R-M compatibility between strains underlying Fig. 5 are found respectively in "5_HGT_RM/HGT_analysis/HGT_events.75_min_bootstrap.tab" and in Supplementary Data 3.

Code availability
The study was performed using a combination of open-source softwares for genomic and multivariate statistical analysis, all cited within the Methods section. In-house scripts used to analyse data are available on Figshare at 10.6084/m9.figshare.24975591.

References
1. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741–746 (2009).
2. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* **6**, 431–440 (2008).
3. Gevers, D. et al. Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3**, 733–739 (2005).
4. Cohan, F. M. Bacterial species and speciation. *Syst. Biol.* **50**, 513–524 (2001).
5. Polz, M. F., Alm, E. J. & Hanage, W. P. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **29**, 170–175 (2013).
6. Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial speciation. *Science* **315**, 476–480 (2007).
7. Didelot, X. & Maiden, M. C. J. Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315–322 (2010).
8. Shapiro, B. J. & Polz, M. F. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol.* **22**, 235–247 (2014).
9. Pigliucci, M. & Muller, G. B. *Evolution, the Extended Synthesis*. (MIT Press, 2010).
10. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
11. Comandatore, F. et al. Gene composition as a potential barrier to large recombinations in the bacterial pathogen Klebsiella pneumoniae. *Genome Biol. Evol.* **11**, 3240–3251 (2019).
12. Majewski, J. & Cohan, F. M. DNA sequence similarity requirements for interspecific recombination in bacillus. *Genetics* **153**, 1525–1533 (1999).
13. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
14. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. USA* **113**, 5658–5663 (2016).
15. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
16. Land, M. et al. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015).
17. Doolittle, W. F. Speciation without species: a final word. *Philos. Theory Pract. Biol.* **11**, https://doi.org/10.3998/ptpbio.16039257.0011.014 (2018).
18. Friman, M. J., Eklund, M. H., Pitkälä, A. H., Rajala-Schultz, P. J. & Rantala, M. H. J. Description of two Serratia marcescens associated mastitis outbreaks in Finnish dairy farms and a review of literature. *Acta Vet. Scand.* **61**, 54 (2019).
19. Dupriez, F., Rejasse, A., Rios, A., Lefebvre, T. & Nielsen-LeRoux, C. Impact and persistence of serratia marcescens in tenebrio molitor larvae and feed under optimal and stressed mass rearing conditions. *Insects* **13**, 458 (2022).
20. Devi, K. A., Pandey, P. & Sharma, G. D. Plant growth-promoting endophyte serratia marcescens AL2-16 enhances the growth of Achyranthes aspera L., a medicinal plant. *HAYATI J. Biosci.* **23**, 173–180 (2016).
21. Chen, Y. P. et al. Phosphate solubilizing bacteria from subtropical soil and their tricalcium phosphate solubilizing abilities. *Appl. Soil Ecol.* **34**, 33–41 (2006).
22. Friedrich, I., Bodenberger, B., Neubauer, H., Hertel, R. & Daniel, R. Down in the pond: Isolation and characterization of a new Serratia marcescens strain (LVF3) from the surface water near frog's lettuce (Groenlandia densa). *PLoS ONE* **16**, e0259673 (2021).
23. Petersen, L. M. & Tisa, L. S. Friend or foe? A review of the mechanisms that drive Serratia towards diverse lifestyles. *Can. J. Microbiol.* **59**, 627–640 (2013).
24. Selvakumar, G. et al. Cold tolerance and plant growth promotion potential of Serratia marcescens strain SRM (MTCC 8708) isolated from flowers of summer squash (Cucurbita pepo). *Lett. Appl. Microbiol.* **46**, 171–175 (2008).
25. Abreo, E. & Altier, N. Pangenome of Serratia marcescens strains from nosocomial and environmental origins reveals different populations and the links between them. *Sci. Rep.* **9**, 46 (2019).
26. Saralegui, C. et al. Genomics of Serratia marcescens isolates causing outbreaks in the same pediatric unit 47 years apart: position in an updated phylogeny of the species. *Front. Microbiol.* **11**, 451 (2020).
27. Ono, T. et al. Global population structure of the Serratia marcescens complex and identification of hospital-adapted lineages in the complex. *Microb. Genomics* **8**, 000793 (2022).
28. Matteoli, F. P., Pedrosa-Silva, F., Dutra-Silva, L. & Giachini, A. J. The global population structure and beta-lactamase repertoire of the opportunistic pathogen Serratia marcescens. *Genomics* **113**, 3523–3532 (2021).
29. Williams, D. J. et al. The genus Serratia revisited by genomics. *Nat. Commun.* **13**, 5195 (2022).
30. Alvaro, A. et al. Cultivation and sequencing-free protocol for Serratia marcescens detection and typing. *iScience* **27**, 109402 (2024).

31. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

32. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

33. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *N. Microbes N. Infect.* **7**, 72–85 (2015).

34. Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2022).

35. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **5**, 86894 (2014).

36. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2008).

37. Moxon, E. R. & Kroll, J. S. The role of bacterial polysaccharide capsules as virulence factors. In *Bacterial Capsules. Current Topics in Microbiology and Immunology*, (eds. Jann, K. & Jann, B.) vol 150, (Springer, Berlin, Heidelberg, 1990).

38. Fernebro, J. et al. Capsular expression in Streptococcus pneumoniae negatively affects spontaneous and antibiotic-induced lysis and contributes to antibiotic tolerance. *J. Infect. Dis.* **189**, 328–338 (2004).

39. Chewapreecha, C. et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).

40. Croucher, N. J. et al. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).

41. Nucci, A., Rocha, E. P. C. & Rendueles, O. Adaptation to novel spatially-structured environments is driven by the capsule and alters virulence-associated traits. *Nat. Commun.* **13**, 4751 (2022).

42. Montagnani, C. et al. Serratia marcescens outbreak in a neonatal intensive care unit: crucial role of implementing hand hygiene among external consultants. *BMC Infect. Dis.* **15**, 11 (2015).

43. Bayramoglu, G. et al. Investigation of an outbreak of Serratia marcescens in a neonatal intensive care unit. *J. Microbiol. Immunol. Infect.* **44**, 111–115 (2011).

44. Moradigaravand, D., Boinett, C. J., Martin, V., Peacock, S. J. & Parkhill, J. Recent independent emergence of multiple multidrug-resistant Serratia marcescens clones within the United Kingdom and Ireland. *Genome Res.* **26**, 1101–1109 (2016).

45. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).

46. Guédin, S. et al. Novel topological features of FhaC, the outer membrane transporter involved in the secretion of the Bordetella pertussis filamentous hemagglutinin. *J. Biol. Chem.* **275**, 30202–30210 (2000).

47. Doman-Pytka, M., Renault, P. & Bardowski, J. Gene-cassette for adaptation of Lactococcus lactis to a plant environment. *Lait* **84**, 33–37 (2004).

48. Rocha, E. P. C. & Bikard, D. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS Biol.* **20**, e3001514 (2022).

49. Chen, Z. et al. A type I restriction modification system influences genomic evolution driven by horizontal gene transfer in Paenibacillus polymyxa. *Front. Microbiol.* **12**, 709571 (2021).

50. Huo, W., Adams, H. M., Trejo, C., Badia, R. & Palmer, K. L. A Type I restriction-modification system associated with enterococcus faecium subspecies separation. *Appl. Environ. Microbiol.* **85**, e02174-18 (2019).

51. Chen, K. et al. The type I restriction enzymes as barriers to horizontal gene transfer: determination of the DNA target sequences recognised by livestock-associated methicillin-resistant staphylococcus aureus clonal complexes 133/ST771 and 398. *Adv. Exp. Med. Biol.* **915**, 81–97 (2016).

52. Waldron, D. E. & Lindsay, J. A. Sau1: a novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into Staphylococcus aureus and between S. aureus isolates of different lineages. *J. Bacteriol.* **188**, 5578–5585 (2006).

53. Sitaraman, R. The role of DNA restriction-modification systems in the biology of Bacillus anthracis. *Front. Microbiol.* **7**, 164773 (2016).

54. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes de novo assembler. *Curr. Protoc. Bioinform.* **70**, e102 (2020).

55. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).

56. Hoaglin, D. C., Iglewicz, B. & Tukey, J. W. Performance of some resistant rules for outlier labeling. *J. Am. Stat. Assoc.* **81**, 991–999 (1986).

57. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).

58. Gona, F. et al. Comparison of core-genome MLST, coreSNP and PFGE methods for Klebsiella pneumoniae cluster analysis. *Microb. Genom.* **6**, e000347 (2020).

59. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

60. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2009).

61. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

62. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).

63. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).

64. Sharpe, D. Chi-square test is statistically significant: now what? *Practical assessment. Res. Eval.* **20**, 8 (2019).

65. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

66. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

67. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).

68. Tonkin-Hill, G. et al. Robust analysis of prokaryotic pangenome gene gain and loss rates with Panstripe. *Genome Res.* **33**, 129–140 (2023).

69. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

70. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

71. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).

72. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

73. Darriba, D. et al. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* **37**, 291–294 (2019).

74. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

75. Boc, A., Diallo, A. B. & Makarenkov, V. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res.* **40**, W573–W579 (2012).

76. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating networks. *Proc. Int. AAAI Conf. Web Soc. Media* **3**, 361–362 (2009).

77. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2014).

78. Sterzi, L. et al. Additional data for the manuscript "Genetic barriers more than environmental associations explain Serratia marcescens population structure". Data sets. *figshare* https://doi.org/10.6084/m9.figshare.24975591 (2024).

Author contributions
L.S. conceived the work, performed the analyses and drafted the manuscript; R.N. performed the analyses; F.D.M. performed the analyses; M.L.F. provided data and material; F.S. provided data and material; A.S. performed the analyses; HA performed the analyses; S.Pap drafted the manuscript; S.Pan drafted the manuscript; S.G.R. provided data and material; G.B.B. performed the analyses; M.C. provided data and material; A.C. provided data and material; P.P. drafted the manuscript; C.F. provided data and material; D.M.C. provided data and material; G.Z. drafted the manuscript; C.B. drafted the manuscript; F.C. conceived the work, performed the analyses, and finalised the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-024-06069-w.

Correspondence and requests for materials should be addressed to Francesco Comandatore.

Reprints and permissions information is available at http://www.nature.com/reprints

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