On the forces shaping diversity and adaptation in the opportunistic pathogen *Serratia marcescens*

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*]

* Corresponding author

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. Laboratory of Microbiology and Virology, 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

Abstract

Serratia marcescens is an opportunistic pathogen able to cause fatal outbreaks in intensive care units. The bacterium shows genomic flexibility and ability to colonise diverse ecological niches, with recent genomic studies proposing the existence of hospital-associated lineages. In this study, we investigated the genetic and ecological drivers underlying the species' evolution. Comparative genomic analyses revealed a population composed of five deeply-demarcated genetic clusters, with independent adaptive trajectories emphasised by profound genetic differences and diverse isolation sources. The clusters also show limited inter-cluster gene flow, partially explained by Restriction-Modification (R-M) systems incompatibility. The clusters emergence dates back to thousands years ago, suggesting that major adaptation events long pre-dated the origin of modern hospitals. Furthermore, metagenomic investigation revealed that the hospital-associated cluster is not enriched in the human gut, possibly having a fresh/marine water reservoir. Overall, this study revisits the diversity of *Serratia marcescens* in light of ecological and genetic barriers.

Introduction

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 sophisticated theoretical framework relies on barriers to recombination to explain the origin and maintenance of divergent sequence clusters[5–8]. 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 genetic barriers and the formation of separate cohesive genotypic clusters. Indeed, recombination rates decrease drastically with sequence divergence[9–11]. 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[6,12,13]. 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)[14,15]. 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 sexual isolation, are mainly focused on "core" genes[16] 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 be also commonly isolated from a multitude of environmental sources, including animal vertebrates[17], insects[18], plants[19], soil[20] and aquatic environments[21,22]. Several evidence of plant-growth promoting activity[19,23] 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[24–27] revealed the existence of a certain number of well-defined clades. The most recent and comprehensive studies[26,27], 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[28] 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 1,113

genomes (see Methods, Table S1 for details)). The Global dataset comprises: i) 230 *S. marcescens* genomes sequenced as part of a large study involving six hospitals in Northern Italy (Alvaro et al. 2023, submitted); 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 Figure 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 (Figures S1-S3). The clusters are coherent with the phylogenetic reconstruction (Figure 1a) and demarcated by deep divisions in the tree: Cluster 1, Cluster 2, Cluster 3 and Cluster 5 are highly-supported (bootstrap supports 100) monophyletic groups, while

Cluster 4 is paraphyletic. The relationships between Cluster 2, Cluster 3 and Cluster 5 are not fully resolved (i.e. bootstrap support < 75).

Despite Cluster 1 comprises 53% (475/902) of the strains within the Global genomic dataset, the cumulative distribution of intra-cluster SNP distance (Figure S4) and Mash distance (Figure S5) show that Cluster 4 and Cluster 5 have higher genetic variability than the other clusters. The Average Nucleotide Identity (ANI) between strains of different clusters, approximated from their Mash distances[29], draws near (and in some cases exceeds) the 95% ANI-based species boundary[30]. Indeed, the maximum ANI among the clusters range from the 96.78% for Cluster 1 - Cluster 4 pair to the 95.56% of Cluster 3 - Cluster 4 (Figure 1b).

Overall, the population structure of *Serratia marcescens* reveals the existence of well-differentiated genetic clusters with clear genetic boundaries, suggesting a remarkable intraspecies genetic diversity.

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, Chi-squared test has revealed an uneven distribution of the clusters in the global continents (X-squared = 103.11, df = 20, p-value = 3.486e-13) and the analysis of the residuals showed that Cluster 5 is associated with North America and negatively associated with Europe (Figure S6). 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 (Figure S7).

The geographically-balanced analysis of the association between Cluster and isolation source (see Methods) indicates that Cluster 1 is significantly associated with clinical settings and negatively associated with environmental sources (Figure 1c). 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).

In 2018, Abreo and Altier[24] 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[26,27]). 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.

Specific genomic features highlight diversity between clusters

Genomic features, such as genome size and GC content, were compared between the five clusters (Figure S8). 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 ($p = 1.8e-05$, Mann-Whitney U test), Cluster 2 ($p = 1.4e-07$), Cluster 3 ($p = 2.5e-10$) and Cluster 5 ($p = 2.2e-06$). Cluster 1 genomes are also significantly larger than genomes in Cluster 3 ($p = 0.00013$, Mann-Whitney U test). Despite all *S. marcescens* strains display 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 ($p < 2.22e-16$, Mann-Whitney U test), Cluster 3 ($p < 2.22e-16$), Cluster 4 ($p < 2.22e-16$) and Cluster 5 ($p = 3.8e-14$). At the same time, Cluster 2 also has a lower GC content than Cluster 3 ($p < 2.22e-16$, Mann-Whitney U test), Cluster 4 ($p = 4.3e-06$) and Cluster 5 ($p = 3.8e-14$). These results are coherent with those recently published by Ono and colleagues[26], in which the authors proposed that the difference in GC content could be a consequence of Horizontal Gene Transfer (HGT).

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 (Figure S9).

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

Specific gene repertoires suggest clusters ecological adaptations

The *S. marcescens* pan-genome comprises a total of 57,700 genes: 2,811 core genes (present in >=95% strains), 3,286 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, Figure S10). On the other hand, the five clusters show pan-genomes of different size (Figure 2a): 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[31].

Interestingly, the PCoA on gene presence absence (Figure 2b) clearly groups the strains coherently with the phylogenetic clusters, while the K-means unsupervised clustering separated Cluster 1 and Cluster 3, and grouped together Clusters 2, 4 and 5 (Figures S1-S3). This result shows that the five *S. marcescens* clusters have specific gene content, especially Cluster 1 and Cluster 3, previously found to be associated with clinical and environmental sources, respectively.

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 (Figure 2c). Among the genes associated to Cluster 1 (see Table S2) there are *fha*C, with a role in the haemolytic process, *fim*C and *hrp*A, both involved in fimbrial biogenesis and *lac*Z, linked to coexistence with mammals. The genes associated with Cluster 3 included genes involved in the metabolism of plant and fungal carbohydrates (*pul*E, *pul*C, *pul*K and *pul*L). Perhaps, these traits could be linked to the differential ecology of the clusters and offer a genetic hint of their adaptation process. Indeed, *fha*C is essential for pathogenicity in *Bordetella pertussis*[32] and the acquisition of lacZ, as stated above, can be linked to the association with mammalian hosts[33]. The presence in Cluster 3 of four genes from the *pul* operon could provide a signal of adaptation to plant environments[34].

In summary, not only that clusters have a gene content of different size, coherently with the intra-cluster genetic variability (see above), but also that Cluster 1 and Cluster 3

have a unique gene repertoire. Since these clusters were notably found to be enriched in clinical and environmental samples, their gene repertoire is coherent with their independent adaptive trajectories towards cluster-specific lifestyles.

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 and Figure S11). 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, but the relationships between the clusters were resolved (Figure S12) and all the clusters corresponded to monophyletic highly-supported groups (bootstrap supports 100).

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 r/m ratio of 2.35, being significantly less recombinogenic than what estimated for opportunistic pathogens[35] 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 recombinations were mapped on basal nodes of all the clusters except for Cluster 4, suggesting that all the other clusters emerged in correspondence of major recombination events. Recombination parameters were also 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 Figure S13.

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 (Figure 3). The bacterial capsule is a well-known virulence factor [36,37] and capsular locus have been shown to be recombination hotspots as consequence of immune escaping [38,39]. 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 [40].

Serratia marcescens clusters have non-synonymous basal mutations on several genes

Starting from the whole-genome alignment, genic coreSNPs were categorised as synonymous/non-synonymous and their origin on the phylogenetic tree was inferred (see Methods, Figure S14). Synonymous and non-synonymous mutations occurring in correspondence of the cluster basal nodes were extracted. A total of 476 genes were enriched in non-synonymous mutation on these nodes: 134 genes for the basal nodes of Cluster 1, 108 genes for Cluster 2, 137 for Cluster 3, 13 for Cluster 4 and 84 for Cluster 5 (see Table S3 for details). 56 out of these 476 genes (12%), having a number of non-synonymous mutations > 1% of the gene length, were considered strongly divergent on the relative cluster basal node. Among these 56 genes, 36 (64%) are related to Cluster 3, 13 (23%) to Cluster 1, five (9%) to Cluster 2, one (2%) to Cluster 5 and one (2%) to Cluster 4. Divergent genes in Cluster 3 are mainly involved in pathways relative to the relationship with environmental settings or other organisms (e.g. membrane proteins, porin proteins and a chitin binding protein), while Cluster 1 genes include four fimbrial genes and three putative membrane genes. These results highlight how the interaction with other organisms (competitors or hosts) could have deeply affected the evolution of *S. marcescens*.

Molecular clock analysis reveals that clusters diverged anciently

Among the 1,062 core genes extracted from the whole-genome alignment, 455 showed signal of HGT (see section “Clusters exhibit partial sexual isolation and preferential gene flow routes” below), 636 resulted to be recombined and one had a ratio of Ka

(number of nonsynonymous substitutions per nonsynonymous site) on Ks (number of synonymous substitutions per synonymous site) above 1 (for details see Table S4). The 279 genes found to be free of HGT, recombination or positive selective pressure were concatenated and subjected to molecular clock analysis. The obtained tree is reported in Figure 4 and the dates estimated for the emergence of the clusters are: Cluster 1 tMRCA median 11,932 ya (95% HPD 1,933 ya - 53,859 ya), Cluster 2 10,046 ya (1,313 ya - 47,042 ya), Cluster 3 11,873 ya (2,203 ya - 54,879 ya), Cluster 4 median 29,384 ya (5,138 ya - 135,900 ya), Cluster 5 median 21,058 ya (2,820 ya - 99,007 ya).

Despite the aim of the analysis was to date the origin of the five clusters, the accurate age estimation of basal nodes is challenged by the intrinsic uncertainty in dating deep ancient nodes with a set of contemporary samples[41]; nevertheless, these results allow some biological considerations. The five *S. marcescens* clusters are ancient and emerged thousands of years ago. As depicted by the distribution of age estimations obtained from the software BEAST (Figure S15), Cluster 1, Cluster 2 and Cluster 3 have likely emerged in the same period, around 10,000 years ago, while Cluster 4 and Cluster 5 are more difficult to date exactly but could be more ancient. This result is consistent with the idea that the five clusters represent distinguished clonal subpopulations and highlights their stability over time, reflecting the high level of differentiation observed between clusters. Interestingly, age estimations suggest that Cluster 1, Cluster 2 and Cluster 3 emerged at the end of the Last Glacial Period, in correspondence to the Neolithic Revolution, with the development of stable human settlements and agriculture.

Clusters exhibit partial sexual isolation and preferential gene flow routes

We established that the *S. marcescens* population is divided in five well-delimited and long-lasting clusters, emerged also by large recombination 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 returned every HGT event inferred by comparing species tree and gene tree. Out of 676 HGT events inferred from 1,062 core genes, 517 (76.4%) occurred among strains of the same Cluster: 111 out of these 517 (21%) occurred within Cluster 1, 118 (23%) within Cluster 2, 61 (12%) within Cluster 3, 113 (22%) Cluster 4 and 114 (22%) within Cluster 5 (Figure 5a). Moreover, inter-cluster HGTs occur preferentially between certain pairs: among the 159 inter-cluster HGTs, 69 (43.4%) occur between Cluster 2 and Cluster 4, and 31 (19.5%) between Cluster 3 and Cluster 5. As shown in Figure 5b, the residuals of the Chi-squared test clearly support that HGT events mainly occur between strains of the same cluster. Interestingly, Cluster 1 (associated with clinical samples and a unique gene repertoire) is the most sexually isolated from 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 a sexual isolation 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, Figure S16). 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 graph, connecting the strains if they harbour compatible R-M systems, was produced. The obtained R-M graph is reported in Figure S16. As shown in Figure 5c, the analysis of the preferential association between cluster pairs (assessed by Chi-squared Test) showed that the strains in Cluster 1, Cluster 2 and Cluster 5 have a strong intracluster preferential R-M compatibility.

As shown in Figure 5d, the Chi-square Test residuals obtained by the R-M systems resulted significantly correlated to those obtained by HGT analysis (Linear regression, $p = 0.0008$). This result suggests that the observed sexual isolation of the clusters could be partially explained by genetic barriers due to R-M system incompatibility.

The ecology 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 et al. 2023 (submitted) 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 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 S17). This suggests that the peculiar 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 shaped by extensive interspecies gene flow[28].

To study the distribution of the *S. marcescens* clusters the protein markers were searched into MGnify[42], a huge 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 6,476 metagenomic-based assemblies possibly containing *Serratia marcescens* sequences. Among these *S. marcescens*-positive assemblies, 5,440 (84%) resulted positive to at least one *S. marcescens* cluster, and 1,720 (32%) resulted positive to a single cluster. Despite a general biome co-presence was observed for the clusters (Figure S18), some interesting statistically significant associations were observed: Cluster 1 with aquatic biomes (i.e. marine and freshwater), Cluster 2 with digestive system and Cluster 5 with plastic surfaces (terephthalate) (For details see Figure 6).

This analytical approach presents some shortcomings. 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[43]. Nevertheless, this analysis represents a useful tool to broaden our knowledge on the ecology of *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.

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. 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 virulence/resistance traits[24–28].

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 show the existence of five well-defined clusters that exhibit remarkable genetic differences, including specific core genes, non-synonymous basal mutations and recombination patterns. Cluster 1 is frequently associated with hospital settings: it has one of the most reduced pan-genome sizes and it comprises a very specific gene repertoire, including genes involved in virulence mechanisms. Furthermore, this cluster has a large number of conserved non-synonymous mutations on the basal node, in particular on fimbrial genes. Fimbrial genes are known pathogenicity factors and a recent study on the *Serratia* genus also highlighted the presence of a large lineage enriched in fimbrial genes[28]. This cluster also has limited intra-cluster genomic variability and 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. 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 is not associated with any isolation source and does not have a highly peculiar gene composition, but metagenomic analysis suggests its association with 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. This cluster presents a high number of conserved non-synonymous basal SNPs and a reduced genetic variability, traits coherent with a process of niche specialisation. Furthermore, this cluster 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, but very few specific genes and few conserved non-synonymous basal mutations. Cluster 5 is enriched in environmental and animal sources while investigation in metagenomic data suggests its ability to survive on plastic surfaces. It contains a large genetic variability and it is the only cluster with a macro geographic association (North America).

The existence of deeply demarcated clusters in *S. marcescens* suggests that geographical, ecological or genetic barriers could maintain this diversity, but the analysis of isolation sites excludes clear geographical barriers. Coherently, genetically distant strains were often isolated in the same hospital, in our dataset as well as in the study published by Moradigaravand and colleagues[44]. Considering the limited geographical isolation, the ecological overlap between the clusters was investigated

with a “reverse ecology” approach, using gene content and basal mutations to trace signals of ecological adaptation[8] . Interestingly, Cluster 1 and Cluster 3 show 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[24], that *S. marcescens* has diverged into an “environmental” clade and a “clinical” clade. Genetic differentiation between the clusters was also investigated 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, also on the basal nodes of the clusters. Despite this, the minimal number of inter-cluster horizontal gene transfer (HGT) events suggests a reduced genetic exchange among the clusters, partially explained by Restriction-Modification (R-M) incompatibility. These two results suggest that genetic barriers could have a major role in the sexual isolation of the clusters, shaping the current population structure of *S. marcescens*. The emergence of the clusters was dated thousands of years ago and Cluster 1, Cluster 2 and Cluster 3 appear to be more recent in comparison to Cluster 4 and Cluster 5. Indeed, Cluster 1, Cluster 2 and Cluster 3 could have emerged simultaneously in correspondence with the end of the Last Glacial Period and the start of the Neolithic revolution, perhaps due to the availability of new niches, similarly to what already proposed for *Campylobacter* species[45]. Moreover, the ancient origin of the clusters suggests that their emergence cannot be explained by adaptation to clinical settings, but as a consequence of ancient global environmental/climate changes.

In conclusion, the genetic landscape portrayed in this study shows five mainly sexually isolated *S. marcescens* clusters emerged thousands of years ago and characterised by

specific gene content. Two clusters in particular show signals of independent adaptive trajectories to clinical and not-clinical environments, respectively. The most clinically relevant cluster has minimal gene flow with other clusters showing a deep separation from the other clusters. Finally, speculating on the origin of the *S. marcescens* cluster, we propose that *S. marcescens* could have separated in allopatric subpopulations during the Last Glacial Period, perhaps due to habitat fragmentation. In this period, these isolated subpopulations independently accumulated specific mutations and gene repertoires, leading to the formation of genetic barriers. Nowadays, all clusters can behave as nosocomial pathogens and often co-occur in the same hospital at the same time, however genetic barriers have become sufficiently thick to maintain the clusters.

Methods

Global genomic dataset preparation

Preliminary genome dataset reconstruction

The preliminary genomic dataset used in the study contained a total of 1,113 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[24,25,44] and absent in the BV-BRC database; iii) 230 *S. marcescens* genome assemblies sequenced in the previous study (Alvaro et al. 2023, submitted); 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 Table S1.

Serratia marcescens genomes sequencing and assembly

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 [46].

Global genomic dataset selection

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 [47] 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[48]: 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" < 5,842 ; iv) "N50" > 7,077 ; v) 5,134 < "ORF count" < 4,594. 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[29] 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[49]: 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

SNP-based phylogenetic analysis

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[50]. The obtained CoreSNPs were subjected to Maximum Likelihood (ML) phylogenetic analysis using RAxML8[51] (with 100 pseudo-bootstraps), using a general time reversible model that incorporates rate variation among sites and proportion of invariant sites (GTR + G + I), according to ModelTest-NG[52]. The obtained tree was manually rooted on the outgroup using Seaview[53]. Lastly, the web-based tool iTOL[54] was used to map strains metadata on the topology.

Clustering analysis

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 - Mash distances) * 100 [29].

For each analysis, the optimal number of clusters was determined in accordance to the best average silhouette score.

Cluster geographic distribution

The geographic distribution of the strains of the different clusters was compared using the Chi-squared 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 [55].

Cluster sampling source enrichment

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 Chi-squared 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 Chi-squared Test of Independence was implemented with a Monte Carlo method: the test was run 1,000 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 1,000 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”.

Clusters comparison

Genomic features

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.

Synteny

The genomic synteny within and between clusters was assessed on the complete genomes available in 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[56] and the intra-cluster synteny plot was obtained using the R package "genoplotR"[57]. The inter-cluster synteny was investigated using one representative strain per cluster.

Pan-genome analysis

Genomes were annotated using Prokka[58] and General Feature Format (GFF) files were fed to Roary[59] 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.[60] Differences in gene content among clusters were also investigated by PCoA on the gene presence/absence Jaccard distance matrix obtained from the Roary tool.

Annotation of cluster-specific core genes

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[61], 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.

Cluster origin and maintenance

Dataset refinement

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 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 0-1,000 SNPs were plotted using R. Herein, this dataset will be referred to as “Refined genomic dataset”.

SNP calling, SNP annotation and Maximum Likelihood phylogenetic analysis

The Purple[50] 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[50]. The extracted coreSNPs were then subjected to ML phylogenetic analysis as described above.

Whole-genome recombination analysis

The ML phylogenetic tree and the whole-genome alignment (obtained using Purple[50]) were fed to ClonalframeML[62] 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 (> 50 kbp) were mapped on the phylogenetic tree.

Analysis of ancestral non-synonymous Single Nucleotide Polymorphisms

The marginal ancestral states of each SNP was reconstructed for each node of the SNP-based phylogenetic tree as described in Yang et al.[63], using the "rerootingMethod" function of the "phytools"[64] R library. Then, the coreSNP originated in the ancestral nodes of each cluster were identified and their positions on the reference genome were plotted using R. Moreover, genes with a significant enhancement of non-synonymous SNPs were identified as outliers in the distribution of values, using Tukey's fences to determine thresholds.

Molecular clock

Molecular clock analysis was carried out on the Refined genome dataset using genes without signal of Horizontal Gene Transfer (HGT), recombination or positive selective

pressure. 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. The obtained alignments were then screened for signal of HGT, gene recombination and positive selective pressure. HGTs were inferred using an approach based on tree bipartition dissimilarity[65]. Core gene alignments were 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[66] 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. Recombination analysis was carried out using the pairwise homoplasy index PHI[67] and recombination was assessed for PHI value < 0.05. Positive selective pressure was evaluated using the “kaks” function of the R library “seqinr[68]”. When the ratio between Ka (number of nonsynonymous substitutions per nonsynonymous site) and Ks (number of synonymous substitutions per synonymous site) exceeds 1. Then gene alignments without signal of HGT (as determined above), recombination or positive selective pressure were selected and concatenated.

The TempEst tool was used to assess the temporal signal in the concatenate and one outlier strain (isolated before 1900 CE) was removed. Then, the alignment and isolation year was fed to BEAST[69] and the analysis was performed with the following parameters: uncorrelated gamma relaxed clock, GTR evolutionary model with empirical base frequencies and gamma site heterogeneity, exponential population growth model of

population size. The analysis was performed 100,000,000 times with samples taken every 10,000 steps. 15,000,000 steps were discarded as a burn-in and Tracer[70] was used to evaluate the convergence of the analysis. The representation of the clocked tree produced by Treeannotator[71] was drawn with Figtree (http://tree.bio.ed.ac.uk/software/figtree/) . The distribution of the dates estimated for the emergence of the clusters was further studied using R.

Gene flow analysis

The HGT analysis performed above 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[72]. Lastly, the preferential association among clusters for HGT events was evaluated analysing the residuals of the Chi-Squared Test.

Analysis of Restriction-Modification (R-M) systems compatibility among the clusters

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[73], 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 Chi-squared Test of Independence.

To investigate whether the R-M system could affect the observed gene flow pattern (as determined above), the Chi-square 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.

Investigation of the S. marcescens clusters enrichment in ecological niches using a large shotgun metagenomics database

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[42] 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[74] (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[75] database. After filtering for coverage and e-value as above, hits were categorised on their taxonomy as hits on *S. marcescens* or as and the same set of percentage of sequence identity values was used. 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.* 2023 (submitted) 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.* 2023 (submitted) 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, Chi- Squared Test of Independence was performed (standard residuals were considered significant if their absolute value was greater than the Bonferroni-corrected critical value).

Captions

Figure 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 and the inner circle around the tree indicate the five clusters coherently and independently determined applying K-means clustering on patristic distances, coreSNP distances and Mash distances. The outer circle around the tree indicates the strain isolation source (blank if not traceable). Tree nodes with bootstrap value > 75 are marked by black dots. b) Distribution of Average Nucleotide Identity (ANI) between *S. marcescens* strains for each cluster. The dark blue vertical lines indicate the species identity threshold (95% ANI). c) Association of *S. marcescens* clusters with animal, clinical or environmental isolation sources. To avoid potential biases due to sampling proximity, Chi-squared test was repeated 1,000 times on geographically-balanced subsets. The box plot illustrates the distribution of the Pearson residuals and the red horizontal lines demarcate the thresholds for statistical significance (see Methods): Cluster 1 is associated with clinical samples (>95% of subsets are significant). Cluster 3 is enriched in environmental isolation sources (52% of subsets). Cluster 5 is enriched in both environmental (76% of subsets) and animal (42% of subsets).

Figure 2. *Serratia marcescens* pan-genome and clusters gene repertoires

a) Pan-genome accumulation curves for each cluster of *S. marcescens* strains. The region between 1 and 300 strains is zoomed to allow a better visualisation of curves belonging to smaller clusters. Clusters are coloured according to the key on the right. 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 highlight the three clusters identified by the K-means unsupervised algorithm. c) The heatmap shows the distribution of cluster-specific core genes among *S. marcescens* strains mapped on the species phylogenetic tree. On the left, the Maximum Likelihood (ML) SNP-based phylogenetic tree including the 902 *S. marcescens* strains of the Global genomic dataset; on the right, the heatmap with a column for each cluster-specific core gene: present genes are coloured on the basis of the associated cluster. The legend is reported at the centre of the figure.

Figure 3. Recombinations along the *Serratia marcescens* genome in each cluster

On the left, distribution of cumulative recombination events along the *Serratia marcescens* genome. The analysis has been performed on a multi-genomes alignment of 86 strains selected to be representative of the clusters (genome alignment was obtained using the genome of *S. marcescens* Db11 strain as reference, position on the x-axis refer to this genome assembly). For each cluster, the recombination distribution is shown as the cumulative number of recombinated bases for genomic windows of 5 kbp. A 10-kbp long highly-recombinated region, containing the capsular genes *wza*, *wzb* and *wzc*, is highlighted by the red box. On the right, comparison among the Maximum Likelihood (ML) phylogenetic trees of the 86 strains, obtained using coreSNPs (on the left) and *wza-wzb-wzc* concatenate (on the right). The results highlight that capsular genes are recombination hotspots.

Figure 4. *Serratia marcescens* molecular clock

On top, dated Maximum Likelihood (ML) phylogenetic tree of *Serratia marcescens* obtained using BEAST. The branches of the tree are coloured on the basis of the clusters (keys on the left) and the time scale (years ago) is reported below the tree. On the bottom, three boxes reporting the geological period (in brown), the Last Glacial Period (in blue) and the Neolithic

period (in green). The grey vertical line highlights the transition to the Holocene (in light brown) and Neolithic period, after the end of the Last Glacial Period.

Figure 5. Gene flow and Restriction-Modification (R-M) systems compatibility

a) Stacked bar charts showing the total number of Horizontal Gene Transfer (HGT) events inferred for each cluster from 1,062 core genes. “Intra” indicates HGT events which happened within strains of the same cluster. “Donor” indicates HGT events departing from the cluster to other clusters. “Recipient” indicates HGT events from other clusters. b) Heatmap showing the Chi-squared residuals of *S. marcescens* clusters pairs on the basis of the number of HGT events inferred between them. c) Heatmap showing the Chi-squared residuals of *S. marcescens* clusters pairs on the basis of R-M systems compatibility. Statistically significant associations are marked with an asterisk (*). d) Linear regression analysis between the residuals of the Chi-squared test on the basis of the R-M compatibility, and the residuals of the Chi-squared test on the basis of the number of HGT events inferred from core genes. Statistically significant associations are marked with asterisks (‘∗’, p value <0.05).

Figure 6. Inference of *S. marcescens* clusters in metagenomic samples

a) The stacked bar chart shows the biomes of metagenomic assemblies from the MGnify database in which *S. marcescens* clusters were identified. An 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 and Cluster 5, the assembly was defined as “Other SMA”. b) The heatmap shows the residuals of the Chi-squared 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).

Supplementary captions

Figure S1. K-means clustering on core Single Nucleotide Polymorphisms distances

a) Silhouette plot of the number of clusters on the basis of core Single Nucleotide Polymorphisms (coreSNP) distances between *S. marcescens* strains. The number of clusters with the highest score was chosen as the optimal number of clusters to fit the data in. b) Principal Coordinate Analysis (PCoA) performed on SNP distances, coloured on the basis of the clusters identified by the K-means algorithm.

Figure S2. K-means clustering on Mash distances

a) Silhouette plot of the number of clusters on the basis of Mash distances between *S. marcescens* strains. The number of clusters with the highest score was chosen as the optimal number of clusters to fit the data in. b) Principal Coordinate Analysis (PCoA) performed on Mash distances, coloured on the basis of the clusters identified by the K-means algorithm.

Figure S3. K-means clustering on patristic distances

a) Silhouette plot of the number of clusters on the basis of patristic distances between *S. marcescens* strains. The number of clusters with the highest score was chosen as the optimal number of clusters to fit the data in. b) Principal Coordinate Analysis (PCoA) performed on patristic distances, coloured on the basis of the clusters identified by the K-means algorithm.

Figure S4. Reverse cumulative distribution of intra-cluster SNP distances

The plot represents the reverse cumulative frequency of the pairwise SNP distances between strains within the same cluster.

Figure S5. Reverse cumulative distribution of intra-cluster Mash distances

The plot represents the reverse cumulative frequency of the pairwise Mash distances between strains within the same cluster.

Figure S6. Macrogeographical distribution of *S.marcescens* clusters

a) Pie charts representing the frequency of each cluster in the six continents (starting from the left: on top North America, Europe, Asia; at the bottom South America, Africa, Oceania). The total number of strains isolated in the continent (n) is shown under each pie chart. b) Chi-squared test was used to assess whether *S. marcescens* clusters are associated with geographical continents. The heatmap shows the Pearson residuals of the Chi-squared test and statistically significant associations are marked with an asterisk (*).

Figure S7. Spatiotemporal distribution of *S. marcescens* in Italian hospitals.

Plot showing the spatiotemporal distribution of 235 *S.marcescens* strains sampled from six Italian hospitals.

Figure S8. GC content and genome length in *S. marcescens* clusters

a) Boxplot showing the percentage of GC content in the genomes of each cluster. The p-value of the Kruskal-Wallis test, performed to test the variance between groups, is shown on the bottom left. The groups with significant pairwise differences (Mann-Whitney U test) are connected and the p-value is written on top.

b) Boxplot showing the genome sizes of each cluster. Kruskal-Wallis test and Mann-Whitneu U test are reported as above.

Figure S9. Synteny between *S.marcescens* clusters

Visualisation of the synteny between the 65 complete genomes present in the study. The synteny is shown for strains within each cluster. On the bottom right ("All clusters"), the synteny between clusters is visualised using one genome representative of the most common syntenic variant for each cluster.

Figure S10. Analysis of the pan-genome openness for *S.marcescens* clusters

The plot shows the log-log regression for new genes found in an increasing number of genome sequences within *S.marcescens* clusters. A slope of the regression line < 1 indicates that when more genomes are analysed, the number of new genes decays slowly, thus the pan-genome is open.

Figure S11. Determination of the optimal coreSNP threshold to create the Refined genomic dataset

In order to reduce the size of the dataset and the genetic bias (for recombination, gene flow and molecular clock analyses) the 902 *S. marcescens* genomes were grouped in coreSNP groups on the basis of a threshold in pairwise coreSNP distance. The determination of the threshold was optimised to reduce and balance the dataset as much as possible with a minor impact on the represented genetic variability. a) To evaluate the potential loss of genetic variability, the mean coreSNP distance for each cluster between strains was plotted for every coreSNP threshold between 0 and 1000 b) To evaluate the size reduction of the dataset, the number of

coreSNP groups was plotted for every coreSNP threshold between 0 and 1000. On the basis of both plots, a threshold of 500 coreSNPs was considered optimal.

Figure S12. Unrooted phylogenetic tree of *S. marcescens* strains in the Refined genomic dataset

SNP-based Maximum Likelihood (ML) phylogenetic tree (performed with 100 pseudo-bootstrap) including the 86 *S. marcescens* strains of the Refined genomic dataset. The five clusters correspond to five monophyletic groups. Bootstraps above 90 are represented as black dots on the corresponding node.

Figure S13. Large recombinations and recombination to mutation ratio in *S. marcescens* clusters

a) Plot illustrating large recombination events (>50 kbp) along the genome (on the right) mapped on every node of the phylogenetic tree (on the left). The colours on the phylogenetic tree represent the *S. marcescens* clusters and the dashed red lines indicate the basal node of each cluster. For Cluster 4, the basal node does not correspond to any large recombination. b) Boxplot showing the recombination to mutation ratio (r/m) for the nodes of each cluster. The p-value of the Kruskal-Wallis test, performed to test the variance between groups, is shown on the bottom left. The groups with significant pairwise differences (Mann-Whitney U test) are connected and the corresponding p-value is written on top.

Figure S14. Synonymous and non-synonymous mutations on the *S. marcescens* clusters basal nodes

The plot displays the mutations that have been mapped on the basal node of the *S. marcescens* clusters. Each dot represents the number of mutations occurring on one gene along the

genome: grey dots indicate the number of synonymous mutations; coloured dots indicate the number of non synonymous mutations.

Figure S15. Distribution of age estimations related to the cluster basal nodes

Plot illustrating the distribution of estimated age values for each cluster returned by the Bayesian molecular clock analysis.

Figure S16. Graph of Restriction-Modification (R-M) systems compatibility

The graph of R-M system compatibility between *Serratia marcescens* strains. The nodes represent the 86 strains selected for the analysis: the strains are connected if they harbour compatible R-M systems.

Figure S17. Sequence similarity between cluster-specific core genes and sequences outside *S. marcescens*

Boxplot showing the percentages of nucleotide identity obtained comparing protein sequences of cluster-specific core genes against protein sequences not belonging to *Serratia marcescens* on the NCBI nr database.

Figure S18. Presence of *S. marcescens* clusters in metagenomic samples from different ecological sources/biomes

The presence of Cluster 1, Cluster 2, Cluster 5 and other S. *marcescens* was investigated in metagenomic samples of the MGnify database searching for cluster-specific core genes and *S. marcescens* core genes. The heatmap shows the frequency with which every combination of clusters was found in samples from different ecological sources/biomes. In brackets, the number of samples for each combination.

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**, (2007).
7. Didelot, X. & Maiden, M. C. J. Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315 (2010).
8. Shapiro, B. J. & Polz, M. F. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends Microbiol.* **22**, 235–247 (2014).
9. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
10. 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).

11. Majewski, J. & Cohan, F. M. DNA Sequence Similarity Requirements for Interspecific Recombination in Bacillus. *Genetics* **153**, 1525–1533 (1999).
12. Oliveira, P. H., Touchon, M. & Rocha, E. P. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, (2014).
13. Oliveira, P. H., Touchon, M. & Rocha, E. P. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. U. S. A.* **113**, (2016).
14. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, (2005).
15. Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, (2015).
16. Doolittle, W. F. Speciation without species: A final word. *Philos. Theory Pr. Biol.* **11**, (2019).
17. 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**, (2019).
18. 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).
19. 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 Journal of Biosciences* **23**, 173–173 (2016).

20. Chen, Y. P. *et al.* Phosphate solubilizing bacteria from subtropical soil and their tricalcium phosphate solubilizing abilities. (2006).
21. 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).
22. Petersen, L. M. & Tisa, L. S. Friend or foe? A review of the mechanisms that drive Serratia towards diverse lifestyles. *Can. J. Microbiol.* **59**, (2013).
23. 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).
24. 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**, (2019).
25. 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**, (2020).
26. Ono, T. *et al.* Global population structure of the Serratia marcescens complex and identification of hospital-adapted lineages in the complex. *Microbial Genomics* **8**, (2022).
27. 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**, (2021).

28. Williams, D. J. *et al.* The genus Serratia revisited by genomics. *Nat. Commun.* **13**, 1–18 (2022).
29. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 1–14 (2016).
30. 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**, 1–8 (2018).
31. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections* **7**, 72 (2015).
32. 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**, (2000).
33. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, (2000).
34. Doman-Pytka, M., Renault, P. P. & Bardowski, J. Gene-cassette for adaptation of Lactococcus lactis to a plant environment. *Lait* **84**, 33–37 (2004).
35. Vos, M. & Didelot, X. A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2008).
36. Moxon, E. R. & Kroll, J. S. The Role of Bacterial Polysaccharide Capsules as Virulence Factors. *Bacterial Capsules* 65–85 (1990).
37. 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).

38. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).
39. Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
40. 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**, 1–15 (2022).
41. Menardo, F., Duchêne, S., Brites, D. & Gagneux, S. The molecular clock of Mycobacterium tuberculosis. *PLoS Pathog.* **15**, (2019).
42. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2022).
43. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **5**, 86894 (2014).
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**, (2016).
45. Sheppard, S. K. & Maiden, M. C. J. The evolution of Campylobacter jejuni and Campylobacter coli. *Cold Spring Harb. Perspect. Biol.* **7**, a018119 (2015).
46. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, (2020).
47. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 1–11 (2010).

48. Hoaglin, D. C., Iglewicz, B. & Tukey, J. W. Performance of Some Resistant Rules for Outlier Labeling. *J. Am. Stat. Assoc.* (2012).
49. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590 (2013).
50. Gona, F. *et al.* Comparison of core-genome MLST, coreSNP and PFGE methods for Klebsiella pneumoniae cluster analysis. *Microbial genomics* **6**, (2020).
51. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312 (2014).
52. 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).
53. 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).
54. 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).
55. Sharpe, D. Chi-Square Test is Statistically Significant: Now What? *Practical Assessment, Research, and Evaluation* **20**, 8 (2019).
56. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS One* **5**, (2010).
57. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
58. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**,

2068–2069 (2014).

59. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
60. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, (2008).
61. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
62. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
63. Yang, Z., Kumar, S. & Nei, M. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650 (1995).
64. Revell, L.J. (2012), phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3: 217-223.
65. Boc, A., Philippe, H. & Makarenkov, V. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst. Biol.* **59**, (2010).
66. 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).
67. Bruen, T. C., Philippe, H. & Bryant, D. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. *Genetics* **172**, 2665 (2006).
68. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. *Structural Approaches to Sequence Evolution* 207–232 (2007).

69. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).

70. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

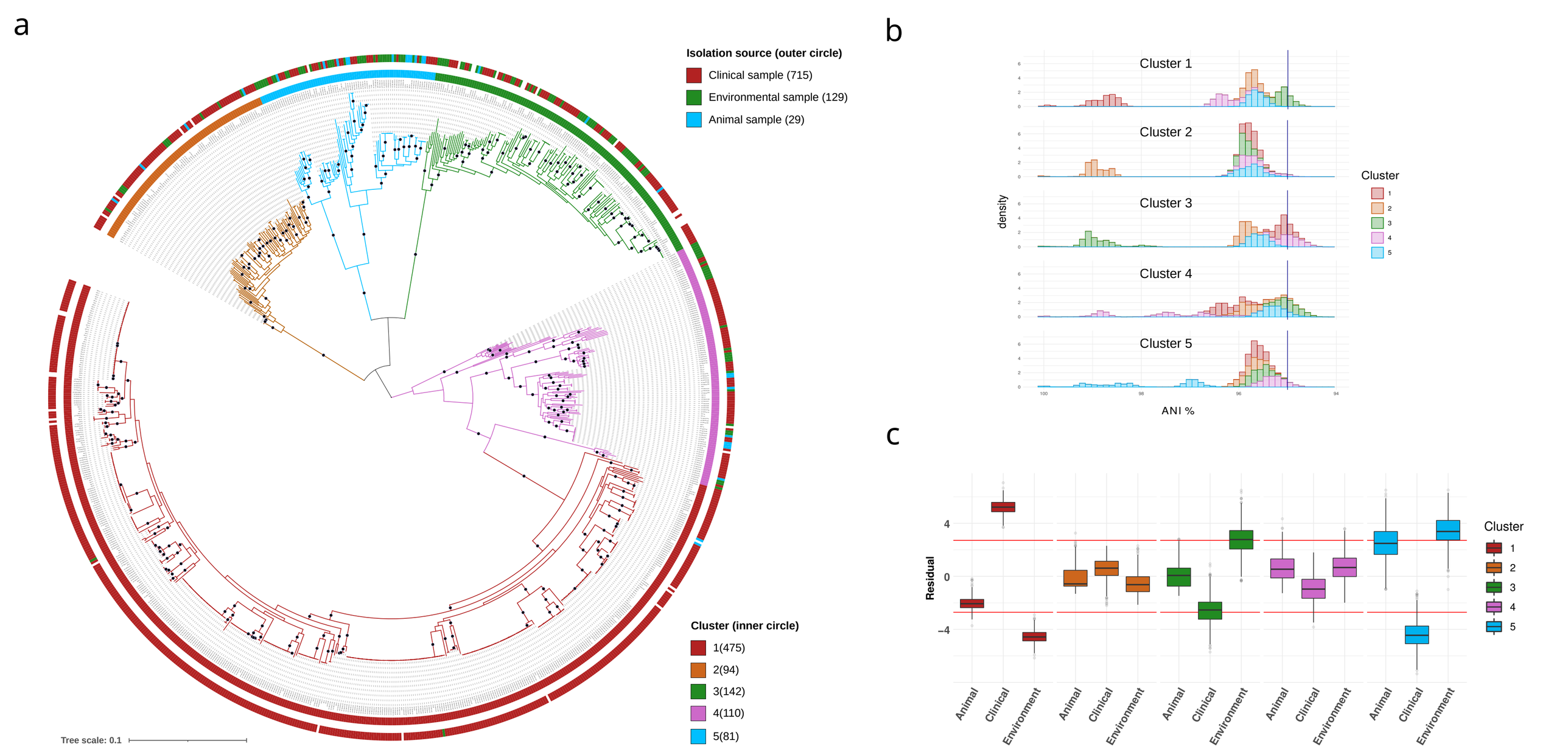
71. Heled, J. & Bouckaert, R. R. Looking for trees in the forest: summary tree from posterior samples. *BMC Evol. Biol.* **13**, 1–11 (2013).

72. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* (2009) doi:10.1609/icwsm.v3i1.13937.

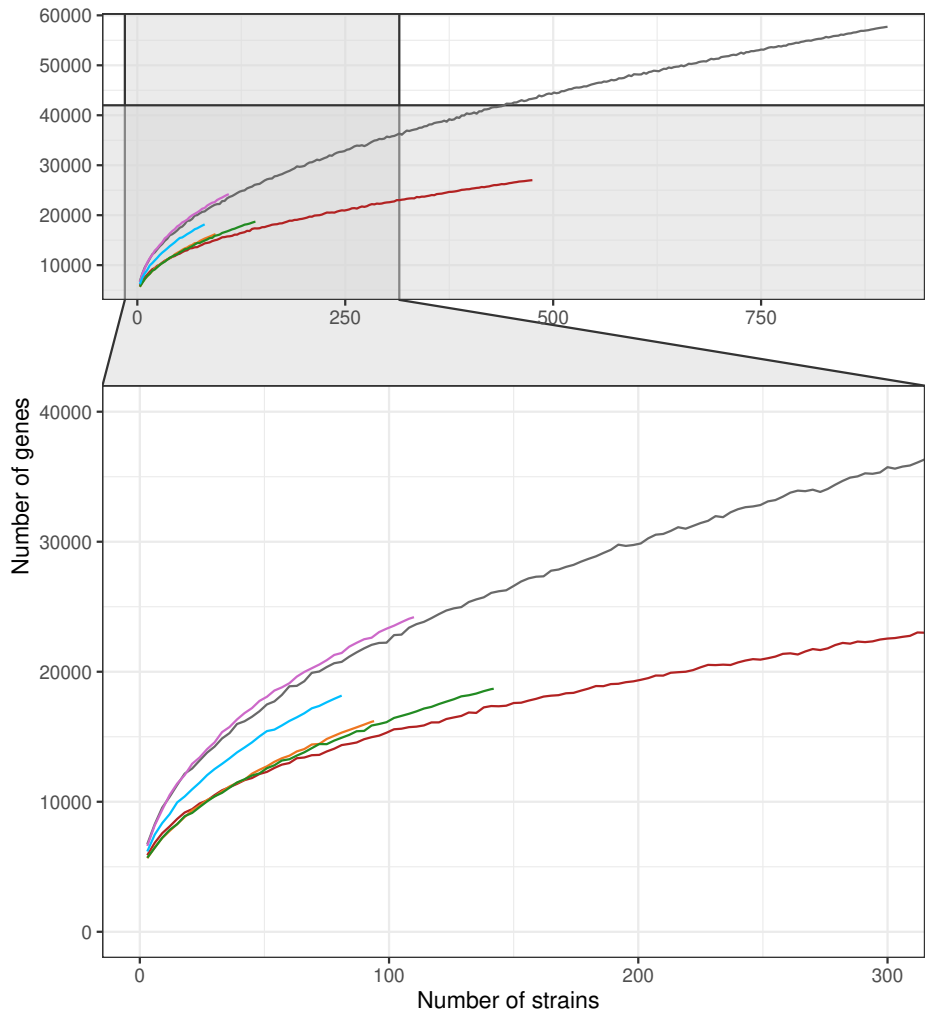
73. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2014).

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

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


a
Isolation source (outer circle)
Clinical sample (715)
Environmental sample (129)
Animal sample (29)
Cluster (inner circle)
1(475)
2(94)
3(142)
4(110)
5(81)
Tree scale: 0.1
b
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
density
ANI %
Cluster
1
2
3
4
5
c
Residual
4
0
−4
Animal
Clinical
Environment
Cluster
1
2
3
4
5

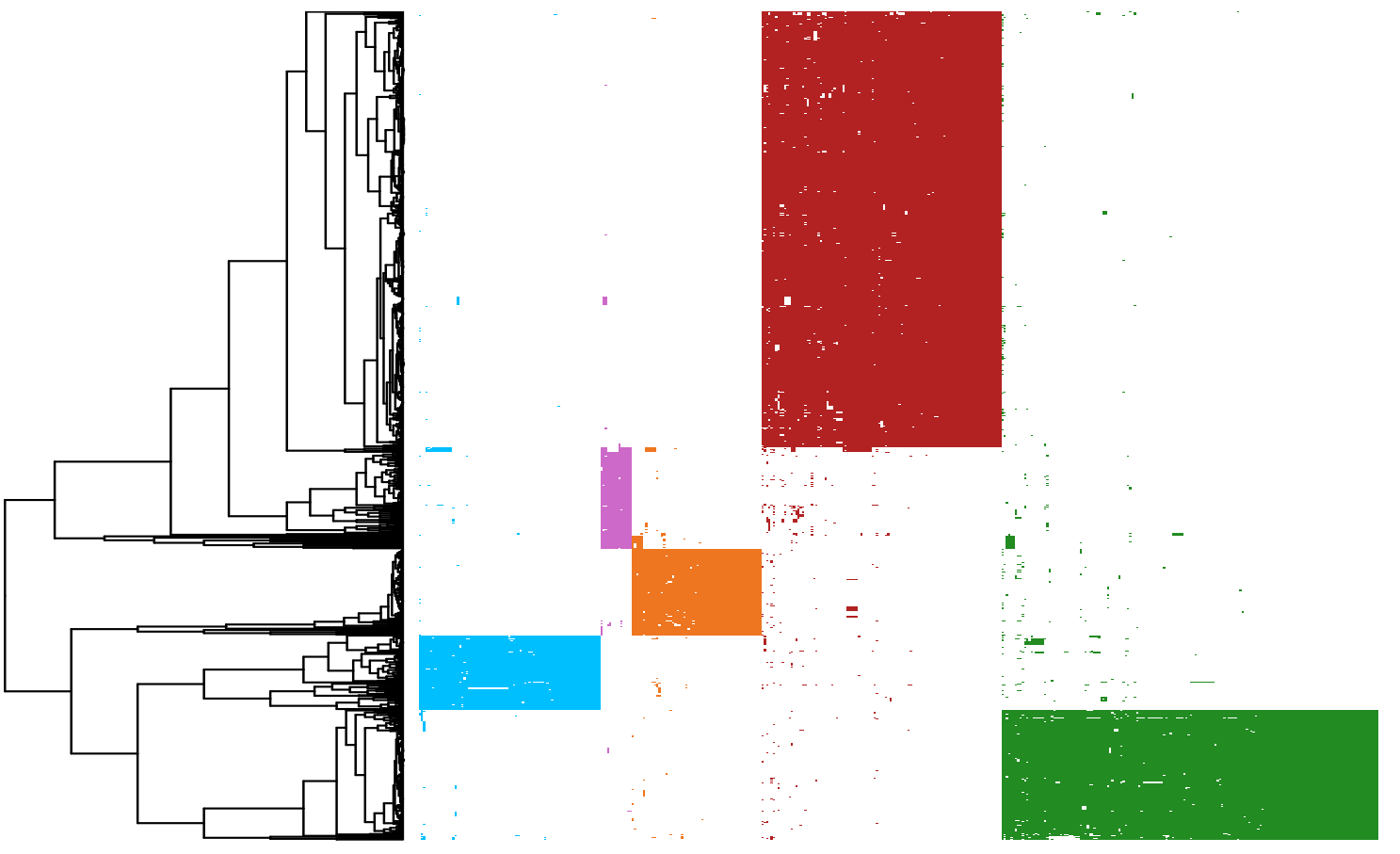
a

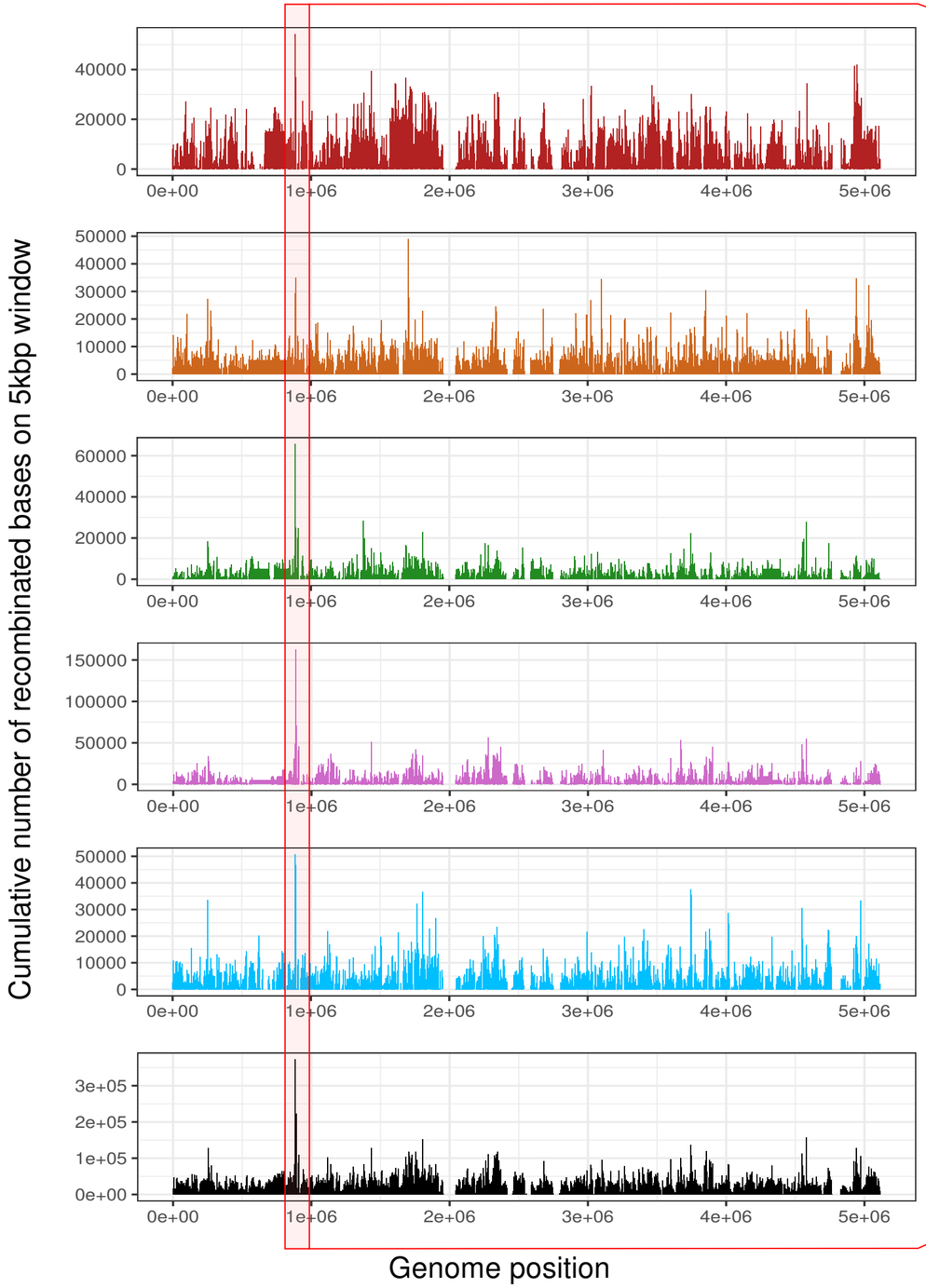

60000
50000
40000
30000
20000
10000
0
250
500
750
Number of genes
40000
30000
20000
10000
0
0
100
200
300
Number of strains

b

Cluster
1
2
3
4
5
All
0.10
0.05
0.00
−0.05
−0.10
Dim2
−0.1
0.0
0.1
Dim1

c

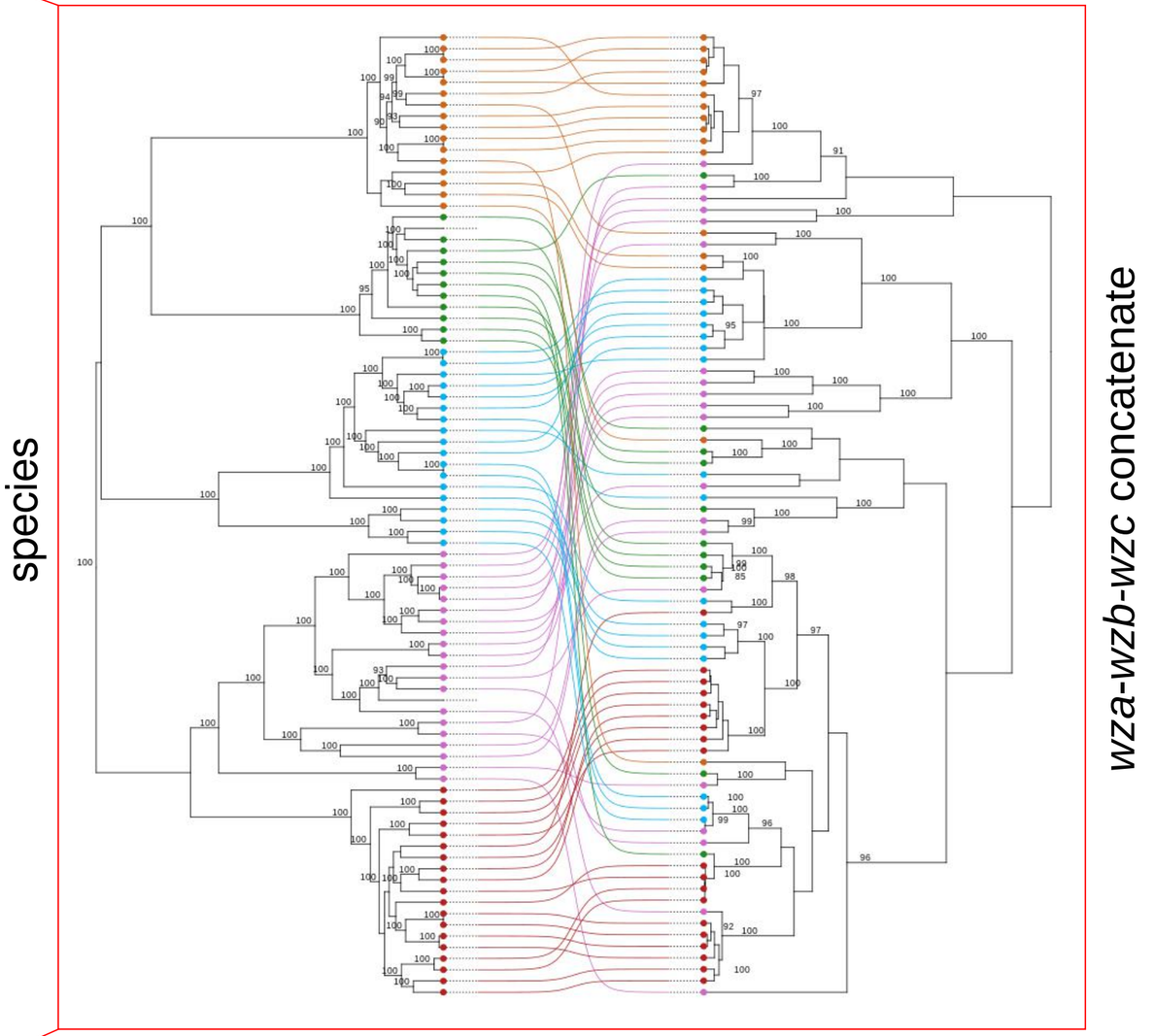


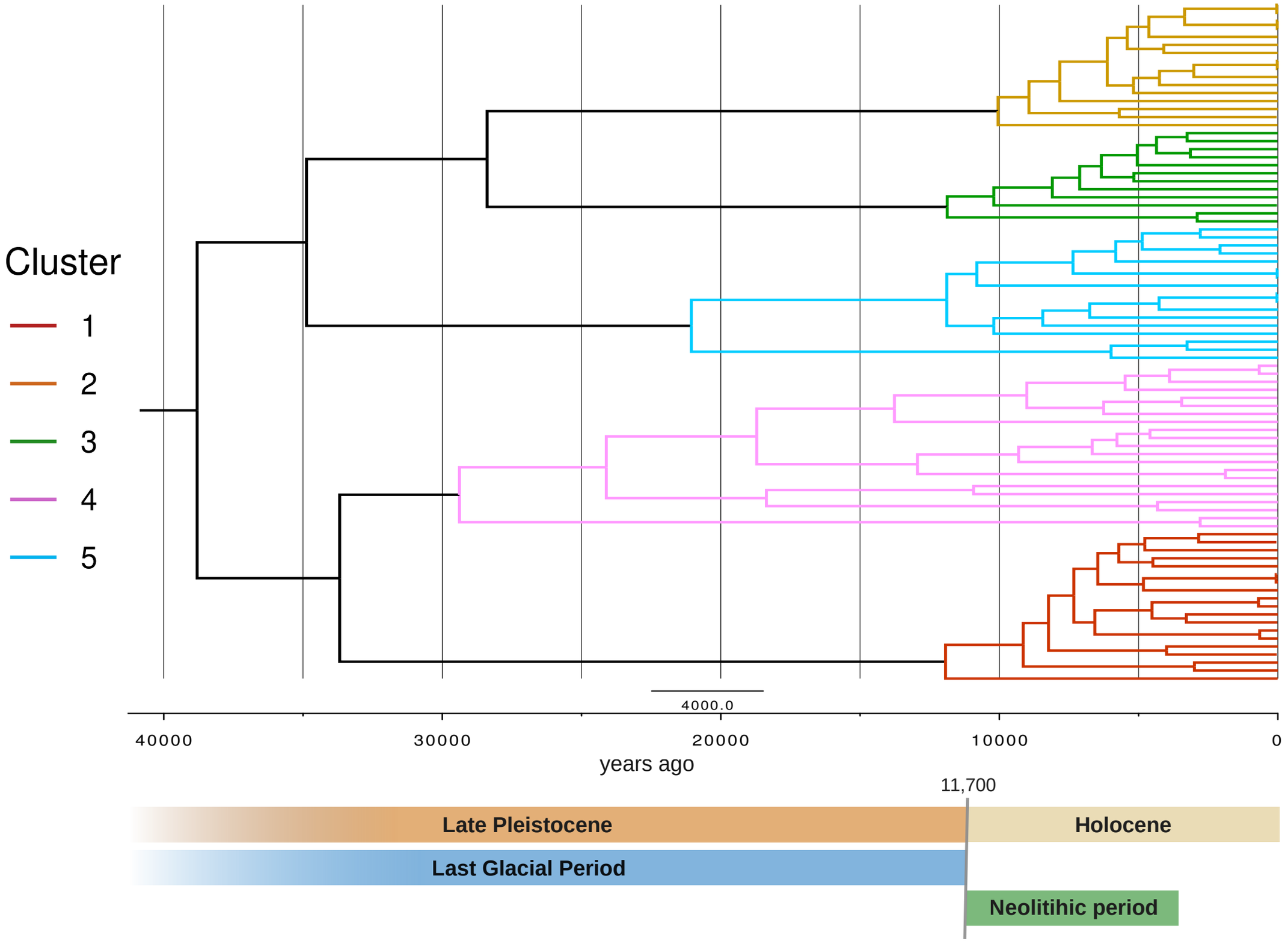


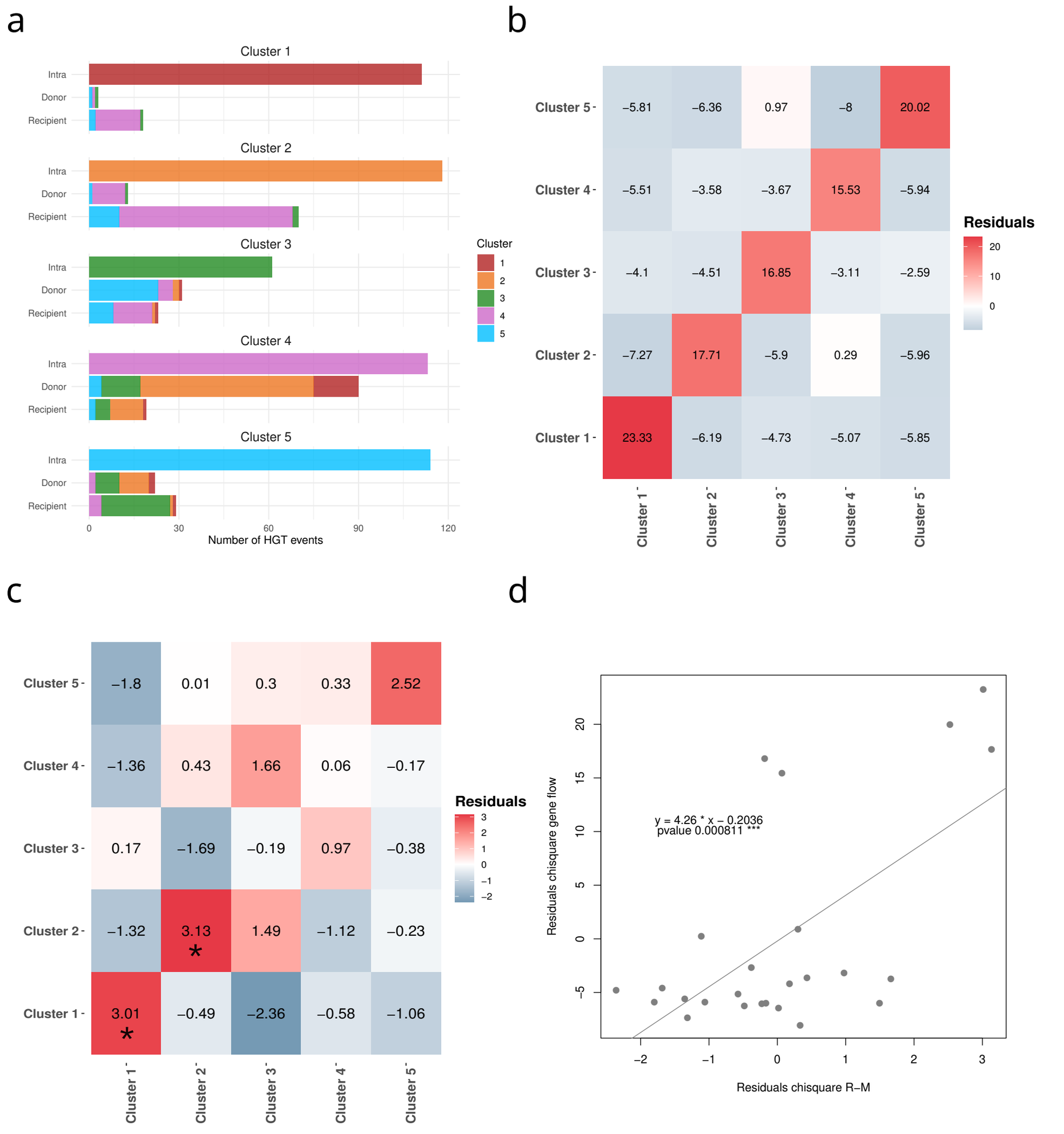
Cluster

- 1
- 2
- 3
- 4
- 5
- All

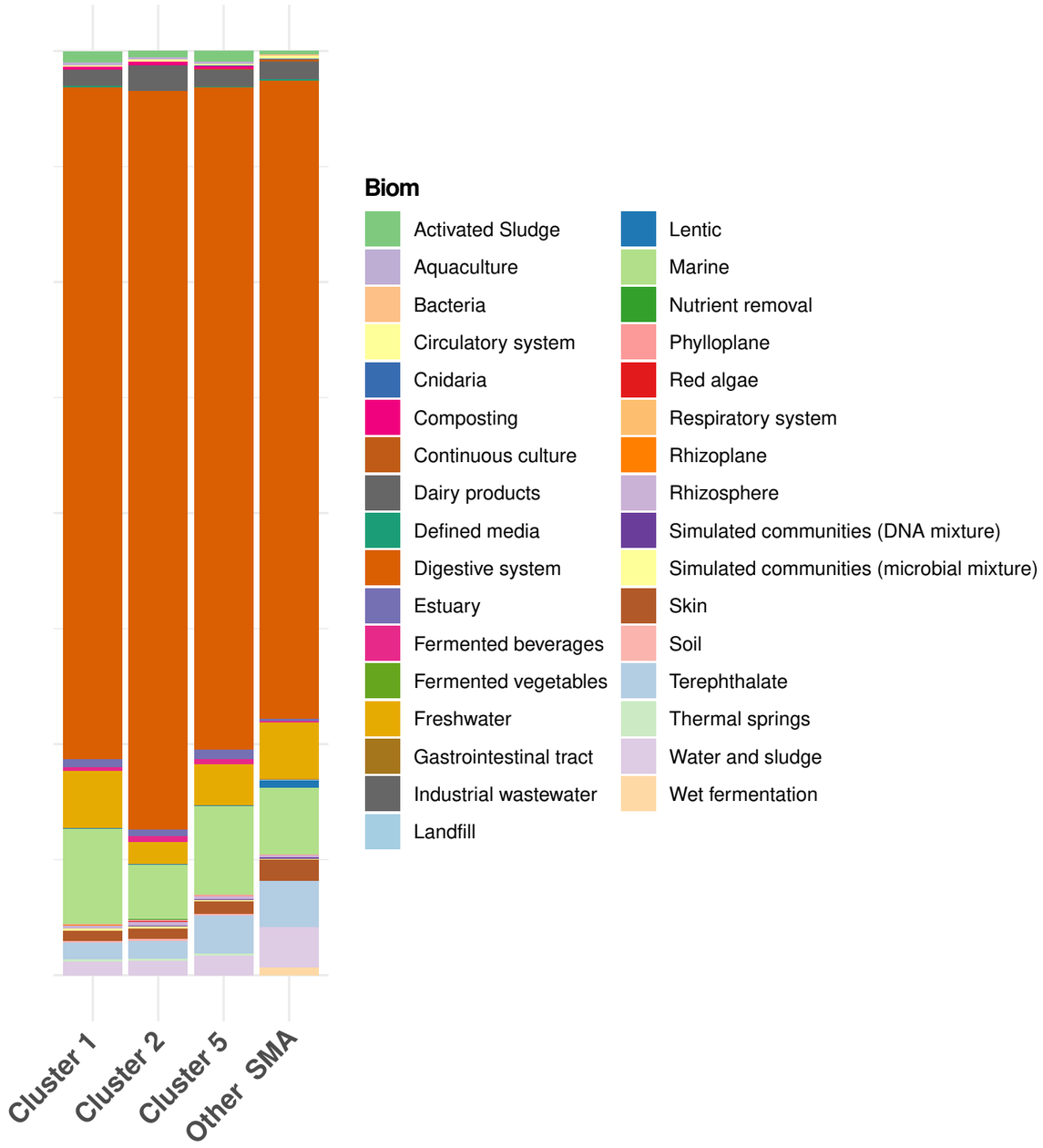
SNP-based vs *wza-wzb-wzc*



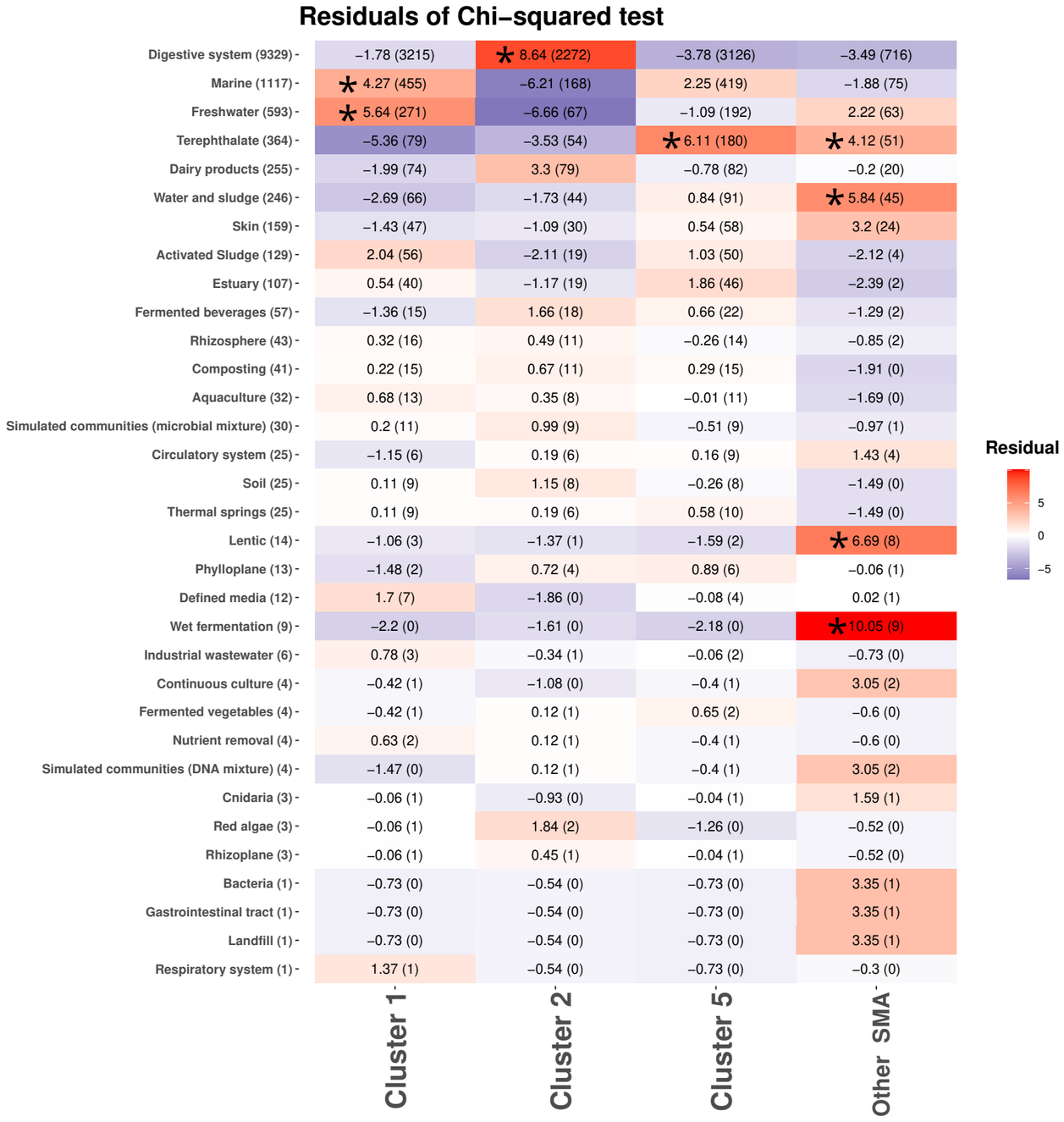

Cluster
1
2
3
4
5
4000.0
40000
30000
20000
10000
0
years ago
11,700
Late Pleistocene
Holocene
Last Glacial Period
Neolitihic period


a
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Intra
Donor
Recipient
0
30
60
90
120
Number of HGT events
Cluster
1
2
3
4
5
b
Cluster 5 -
Cluster 4 -
Cluster 3 -
Cluster 2 -
Cluster 1 -
−5.81 −6.36 0.97 −8 20.02
−5.51 −3.58 −3.67 15.53 −5.94
−4.1 −4.51 16.85 −3.11 −2.59
−7.27 17.71 −5.9 0.29 −5.96
23.33 −6.19 −4.73 −5.07 −5.85
Cluster 1 -
Cluster 2 -
Cluster 3 -
Cluster 4 -
Cluster 5 -
Residuals
20
10
0
c
Cluster 5 -
Cluster 4 -
Cluster 3 -
Cluster 2 -
Cluster 1 -
−1.8 0.01 0.3 0.33 2.52
−1.36 0.43 1.66 0.06 −0.17
0.17 −1.69 −0.19 0.97 −0.38
−1.32 3.13 * 1.49 −1.12 −0.23
3.01 * −0.49 −2.36 −0.58 −1.06
Cluster 1 -
Cluster 2 -
Cluster 3 -
Cluster 4 -
Cluster 5 -
Residuals
3
2
1
0
−1
−2
d
y = 4.26 * x − 0.2036
pvalue 0.000811 ***
Residuals chisquare gene flow
Residuals chisquare R−M
−5
0
5
10
15
20
−2
−1
0
1
2
3

a


Biom
Activated Sludge
Aquaculture
Bacteria
Circulatory system
Cnidaria
Composting
Continuous culture
Dairy products
Defined media
Digestive system
Estuary
Fermented beverages
Fermented vegetables
Freshwater
Gastrointestinal tract
Industrial wastewater
Landfill
Lentic
Marine
Nutrient removal
Phylloplane
Red algae
Respiratory system
Rhizoplane
Rhizosphere
Simulated communities (DNA mixture)
Simulated communities (microbial mixture)
Skin
Soil
Terephthalate
Thermal springs
Water and sludge
Wet fermentation
Cluster 1
Cluster 2
Cluster 5
Other SMA

b


Residuals of Chi−squared test
Digestive system (9329) | −1.78 (3215) | * 8.64 (2272) | −3.78 (3126) | −3.49 (716)
Marine (1117) | * 4.27 (455) | −6.21 (168) | 2.25 (419) | −1.88 (75)
Freshwater (593) | * 5.64 (271) | −6.66 (67) | −1.09 (192) | 2.22 (63)
Terephthalate (364) | −5.36 (79) | −3.53 (54) | * 6.11 (180) | * 4.12 (51)
Dairy products (255) | −1.99 (74) | 3.3 (79) | −0.78 (82) | −0.2 (20)
Water and sludge (246) | −2.69 (66) | −1.73 (44) | 0.84 (91) | * 5.84 (45)
Skin (159) | −1.43 (47) | −1.09 (30) | 0.54 (58) | 3.2 (24)
Activated Sludge (129) | 2.04 (56) | −2.11 (19) | 1.03 (50) | −2.12 (4)
Estuary (107) | 0.54 (40) | −1.17 (19) | 1.86 (46) | −2.39 (2)
Fermented beverages (57) | −1.36 (15) | 1.66 (18) | 0.66 (22) | −1.29 (2)
Rhizosphere (43) | 0.32 (16) | 0.49 (11) | −0.26 (14) | −0.85 (2)
Composting (41) | 0.22 (15) | 0.67 (11) | 0.29 (15) | −1.91 (0)
Aquaculture (32) | 0.68 (13) | 0.35 (8) | −0.01 (11) | −1.69 (0)
Simulated communities (microbial mixture) (30) | 0.2 (11) | 0.99 (9) | −0.51 (9) | −0.97 (1)
Circulatory system (25) | −1.15 (6) | 0.19 (6) | 0.16 (9) | 1.43 (4)
Soil (25) | 0.11 (9) | 1.15 (8) | −0.26 (8) | −1.49 (0)
Thermal springs (25) | 0.11 (9) | 0.19 (6) | 0.58 (10) | −1.49 (0)
Lentic (14) | −1.06 (3) | −1.37 (1) | −1.59 (2) | * 6.69 (8)
Phylloplane (13) | −1.48 (2) | 0.72 (4) | 0.89 (6) | −0.06 (1)
Defined media (12) | 1.7 (7) | −1.86 (0) | −0.08 (4) | 0.02 (1)
Wet fermentation (9) | −2.2 (0) | −1.61 (0) | −2.18 (0) | * 10.05 (9)
Industrial wastewater (6) | 0.78 (3) | −0.34 (1) | −0.06 (2) | −0.73 (0)
Continuous culture (4) | −0.42 (1) | −1.08 (0) | −0.4 (1) | 3.05 (2)
Fermented vegetables (4) | −0.42 (1) | 0.12 (1) | 0.65 (2) | −0.6 (0)
Nutrient removal (4) | 0.63 (2) | 0.12 (1) | −0.4 (1) | −0.6 (0)
Simulated communities (DNA mixture) (4) | −1.47 (0) | 0.12 (1) | −0.4 (1) | 3.05 (2)
Cnidaria (3) | −0.06 (1) | −0.93 (0) | −0.04 (1) | 1.59 (1)
Red algae (3) | −0.06 (1) | 1.84 (2) | −1.26 (0) | −0.52 (0)
Rhizoplane (3) | −0.06 (1) | 0.45 (1) | −0.04 (1) | −0.52 (0)
Bacteria (1) | −0.73 (0) | −0.54 (0) | −0.73 (0) | 3.35 (1)
Gastrointestinal tract (1) | −0.73 (0) | −0.54 (0) | −0.73 (0) | 3.35 (1)
Landfill (1) | −0.73 (0) | −0.54 (0) | −0.73 (0) | 3.35 (1)
Respiratory system (1) | 1.37 (1) | −0.54 (0) | −0.73 (0) | −0.3 (0)
Cluster 1
Cluster 2
Cluster 5
Other SMA
Residual
5
0
−5