

**UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
CURSO DE CIÊNCIAS BIOLÓGICAS**

Letícia Ferreira Machado

**Diversidade genômica, arquitetura do pangenoma e variação funcional no
gênero *Yersinia***

Juiz de Fora

2026

Letícia Ferreira Machado

Diversidade genômica, arquitetura do pangenoma e variação funcional no gênero *Yersinia*

Trabalho de Conclusão de Curso
apresentado ao Instituto de Ciências
Biológicas da Universidade Federal de
Juiz de Fora, como requisito parcial à
obtenção do grau de Bacharel em
Ciências Biológicas.

Orientadora: Prof.^a Dr.^a Julliane Dutra Medeiros

Juiz de Fora

2026

Letícia Ferreira Machado

Trabalho de Conclusão de Curso
apresentado ao Instituto de Ciências
Biológicas da Universidade Federal de Juiz
de Fora, como requisito parcial à obtenção
do grau de Bacharel em Ciências
Biológicas.

Aprovada em 03 de julho de 2026

BANCA EXAMINADORA



Documento assinado digitalmente

JULLIANE DUTRA MEDEIROS

Data: 06/07/2026 13:33:28-0300

Verifique em <https://validar.iti.gov.br>

Prof.^a Dr.^a Julliane Dutra Medeiros -Orientadora
Universidade Federal de Juiz de Fora



Documento assinado digitalmente

MARIANA FONSECA ROSSI

Data: 07/07/2026 12:18:40-0300

Verifique em <https://validar.iti.gov.br>

Prof.^a Dr.^a Mariana Fonseca Rossi
Universidade Federal de Juiz de Fora



Documento assinado digitalmente

LAURA MARIA ANDRADE DE OLIVEIRA

Data: 06/07/2026 13:44:55-0300

Verifique em <https://validar.iti.gov.br>

Dr.^a Laura Maria Andrade de Oliveira
Fundação Oswaldo Cruz

AGRADECIMENTOS

Agradeço ao universo por colocar tantas pessoas boas no meu caminho.

À minha família, em especial aos meus pais, Edson e Rosi, à minha irmã Larissa e à minha avó Maria, por todo o apoio, carinho e dedicação ao longo de toda a minha vida. Vocês são meu alicerce.

Aos meus amigos da graduação e da vida, Liz, Ana Alice, Maria Clara, Gabi, Carol, Giovana, Marina, Gabriel, Cauã, Fernando, Letícia, Maynara e tantos outros, pela companhia, pelo apoio e por terem marcado essa etapa da minha vida da forma mais bonita possível. Obrigada por tanto!

À Maria Clara, Mariana, Laura, Viviane e ao Erick, que me receberam no Laboratório de Imunologia ainda no início da graduação. Com vocês aprendi quase tudo do que sei, a ficar até tarde em experimentos, mas também a sair para confraternizar no dia seguinte. Obrigada por todos os conselhos e pela amizade.

Aos meus amigos do NUPEP, Ludmila, Thaís, José e Fernandas, pela amizade, por terem ensinado e aprendido comigo e, especialmente, por serem sempre um motivo de risada até nos dias mais difíceis.

Aos meus amigos do LABIM, especialmente Laura, Marcelo, Léo e Larissa, por tudo que aprendi com vocês, pela amizade, pelo acolhimento e por terem tornado essa experiência um dos momentos mais especiais da minha graduação.

Às professoras Mariana Rossi e Juciane Castro, que marcaram a minha formação em sala de aula. À Mariana, por me mostrar uma forma mais bonita de ver a Biologia, através do pensamento evolutivo, dando mais sentido à minha formação. À Juciane, por todo o incentivo, pelos conselhos, pela confiança e por sempre acreditar em mim.

Ao Gustavo, por toda a ajuda durante a realização deste trabalho, por compartilhar conhecimentos com tanto entusiasmo, por estar sempre disposto a ajudar e pela amizade. É um prazer aprender com você.

À minha orientadora, Julliane D. Medeiros, pelo acolhimento neste último ano, pela confiança, por todo o incentivo e pela paciência. Que todo aluno tenha a sorte de encontrar orientadores como você, que contribuem de forma tão gentil

para o nosso desenvolvimento. Tem sido um privilégio aprender e trabalhar com você.

À minha professora do ensino médio, Bárbara Muniz, por ter sido o meu primeiro exemplo de bióloga e pesquisadora, e por todo o incentivo. E a todos os demais professores que contribuíram para a minha formação até aqui!

À Universidade Federal de Juiz de Fora, ao RePesq e ao GEMBIO pelo suporte institucional, computacional e científico. E e ao NCBI pela disponibilização dos dados genômicos utilizados neste estudo.

RESUMO

O gênero *Yersinia* compreende espécies amplamente distribuídas em diferentes ambientes e hospedeiros, incluindo patógenos de relevância médica e veterinária, como *Y. pestis*, *Y. enterocolitica* e *Y. pseudotuberculosis*, além de espécies associadas a diversos estilos de vida. Essa diversidade ecológica reflete a ampla variabilidade genômica observada no gênero, particularmente no genoma acessório, o qual pode estar associado à adaptação a diferentes nichos ecológicos. Neste estudo, investigamos a diversidade genômica do gênero *Yersinia* a partir da integração de análises de similaridade nucleotídica média (ANI), reconstrução filogenômica, caracterização do pangenoma, anotação funcional e análise de metadados associados à fonte de isolamento e ao hospedeiro disponíveis em bancos de dados públicos. Para isso, foi utilizado um conjunto de genomas públicos, o qual passou por deduplicação para redução de redundância e viés de amostragem. Foi possível observar um forte viés na distribuição dos genomas depositados, com predominância de *Y. enterocolitica* e *Y. pestis*, seguido por uma expressiva redução do conjunto após a filtragem. A análise baseada em ANI permitiu delimitar *clusters* genômicos compatíveis com espécies previamente descritas, os quais foram corroborados pela estrutura filogenômica. O pangenoma apresentou comportamento aberto, com predominância de genes acessórios em relação ao genoma persistente. Além disso, o conteúdo gênico organizou os genomas em três grandes *clusters*, congruentes com os principais clados evolutivos do gênero. Em termos funcionais, o genoma persistente é majoritariamente representado por categorias relacionadas ao metabolismo e ao processamento de informações celulares, enquanto os genes acessórios apresentaram alta variabilidade, além de uma associação com elementos genéticos móveis. A análise de metadados revelou padrões sutis de distribuição, parcialmente estruturados entre as espécies, embora com ampla sobreposição entre diferentes fontes de isolamento e hospedeiros. Em conjunto, os resultados indicam que a diversidade genômica do gênero *Yersinia* é estruturada em níveis hierárquicos, com concordância entre organização do conteúdo gênico, estrutura filogenética e delimitação taxonômica, além de padrões ecológicos parcialmente consistentes.

Palavras-chave: *Yersinia*; diversidade genômica; evolução bacteriana; pangenoma; variação funcional.

ABSTRACT

The genus *Yersinia* comprises species widely distributed across diverse environments and hosts, including medically and veterinary relevant pathogens such as *Y. pestis*, *Y. enterocolitica*, and *Y. pseudotuberculosis*, as well as species associated with broad lifestyles. This ecological diversity reflects the extensive genomic variability observed within the genus, particularly in the accessory genome, which may be associated with adaptation to different ecological niches. In this study, we investigated the genomic diversity of the genus *Yersinia* through the integration of average nucleotide identity (ANI) analyses, phylogenomic reconstruction, pangenome characterization, functional annotation, and metadata analysis related to isolation sources and hosts available in public databases. To this end, a publicly available genome dataset was used and deduplicated to reduce redundancy and sampling bias. A strong bias in genome distribution was observed, with a predominance of *Y. enterocolitica* and *Y. pestis*, followed by a substantial reduction of the dataset after filtering. ANI-based analysis allowed the delineation of genomic clusters compatible with previously described species, which were further supported by phylogenomic structure. The pangenome exhibited an open behavior, with a predominance of accessory genes over the persistent genome. In addition, gene content organized the genomes into three major clusters, congruent with the main evolutionary clades of the genus. Functionally, the persistent genome was mainly represented by categories related to metabolism and cellular information processing, whereas accessory genes showed higher variability and part was associated with mobile genetic elements. Metadata analysis revealed subtle distribution patterns, partially structured among species and clusters, although with extensive overlap across isolation sources, hosts, and environments. Overall, the results indicate that the genomic diversity of the genus *Yersinia* is structured at hierarchical levels, with agreement between gene content organization, phylogenetic structure, and taxonomic delimitation, as well as partially consistent ecological patterns.

Keywords: *Yersinia*; genomic diversity; pangenome; bacterial evolution; functional variation.

LISTA DE ILUSTRAÇÕES

Figura 1 – Heatmap de identidade nucleotídica média (ANI) entre genomas de *Yersinia*.

Figura 2 – Estrutura filogenômica do gênero *Yersinia*.

Figura 3 – Dinâmica de crescimento do pangenoma de *Yersinia*.

Figura 4 – Estruturação do conteúdo gênico em *Yersinia*.

Figura 5 – Distribuição funcional dos ortólogos do genoma persistente do gênero *Yersinia*.

Figura 6 – Distribuição dos elementos genéticos móveis (MGEs) entre os genes exclusivos das espécies de *Yersinia*.

Figura 7 – Mapeamento dos metadados ecológicos sobre a estrutura do pangenoma de *Yersinia*.

LISTA DE QUADROS

Quadro 1 – Impacto da desreplicação genômica na representatividade das espécies de *Yersinia*.

Quadro 2 – Distribuição das famílias gênicas nas partições do pangenoma de *Yersinia*.

Quadro A1 – Reclassificação taxonômica de genomas do gênero *Yersinia* baseada em similaridade nucleotídica.

LISTA DE ABREVIATURAS E SIGLAS

T3SS – Sistema de Secreção do Tipo III (do inglês *Type III Secretion System*)

HGT – Transferência Horizontal de Genes (do inglês *Horizontal Gene Transfer*)

MGE – Elementos Genéticos Móveis (do inglês *Mobile Genetic Elements*)

NCBI – *National Center for Biotechnology Information*

MAG – Genomas Montados a partir de Metagenomas (do inglês *Metagenome Assembled Genomes*)

ANI – Identidade Nucleotídica Média (do inglês *Average Nucleotide Identity*)

AF – Fração Alinhada (do inglês *Aligned Fraction*)

UFBoot – Ultrarrápido Bootstrap (do inglês *Ultrafast Bootstrap*)

SH-aLRT – teste de razão de verossimilhança aproximado de Shimodaira–Hasegawa (do inglês *Shimodaira–Hasegawa approximate likelihood ratio test*)

PCoA – Análise de Coordenadas Principais (do inglês *Principal Coordinates Analysis*)

COG – Grupos Ortólogos de Genes (do inglês *Clusters of Orthologous Groups*)

SUMÁRIO

1. INTRODUÇÃO.....	13
1.1 O gênero <i>Yersinia</i>	13
1.2 Plasticidade genômica e evolução bacteriana.....	15
1.3 Diversidade ecológica e adaptação bacteriana.....	18
2. HIPÓTESE.....	19
3. OBJETIVOS.....	20
3.1 Objetivo Geral.....	20
3.2 Objetivos Específicos.....	20
4. METODOLOGIA.....	21
4.1 Obtenção e curadoria dos genomas.....	21
4.2 Reconstrução filogenética e delimitação taxonômica do gênero <i>Yersinia</i> ..	22
4.3 Construção e estrutura do pangenoma.....	23
4.4 Anotação funcional.....	24
4.5 Classificação ecológica dos isolados.....	25
5. RESULTADOS.....	26
5.1 Impacto da desreplicação na representatividade genômica de <i>Yersinia</i> ...	27
5.2 Estrutura taxonômica e filogenômica do gênero <i>Yersinia</i>	33
5.3 Estrutura, diversidade e dinâmica do pangenoma de <i>Yersinia</i>	35
5.4 Anotação funcional do genoma persistente e dos genes exclusivos das espécies de <i>Yersinia</i>	39
5.5 Associação entre composição gênica e categorias ecológicas dos isolados de <i>Yersinia</i>	42
6. DISCUSSÃO.....	48
7. CONCLUSÃO.....	49
8. PERSPECTIVAS FUTURAS.....	50
REFERÊNCIAS BIBLIOGRÁFICAS.....	55
APÊNDICES.....	56
APÊNDICE A.....	57

1. INTRODUÇÃO

1.1 O gênero *Yersinia*

O gênero *Yersinia* (família Yersiniaceae, ordem Enterobacterales) compreende 27 espécies descritas de cocobacilos Gram-negativos, não formadores de esporos e anaeróbios facultativos (Le Guern *et al.*, 2020; Rivas *et al.*, 2026). Esses microrganismos apresentam ampla distribuição em ambientes terrestres e aquáticos, sendo frequentemente isolados de solo, água, alimentos e de uma grande diversidade de hospedeiros, incluindo mamíferos, aves, peixes e insetos, refletindo a sua versatilidade ecológica (Sulakvelidze, 2000).

Historicamente, o gênero ganhou destaque em função da relevância clínica de algumas espécies, como a *Yersinia pestis*, responsável pelas formas bubônica, septicêmica e pneumônica da peste (Yang *et al.*, 2023). Entre as espécies associadas a infecções humanas, destacam-se ainda *Y. pseudotuberculosis*, agente da febre escarlatiniforme do Extremo Oriente (Amphlett, 2016) e *Y. enterocolitica*, associada à yersiniose gastrointestinal (Bottone, 1999). Análises filogenômicas sugerem que a patogenicidade surgiu de forma independente nessas linhagens. Enquanto *Y. pestis* e *Y. pseudotuberculosis* formam um mesmo complexo evolutivo, *Y. enterocolitica* pertence a uma linhagem mais divergente dentro do gênero, apresentando biotipos patogênicos (biotipos 1B, 2 a 5) e ambientais (biotipo 1A); no entanto, essa categorização é controversa, uma vez que isolados do biotipo 1A têm sido recuperados de casos sintomáticos (Reuter *et al.*, 2014; Alenizi *et al.*, 2016). A evolução da patogenicidade nestes clados envolveu, de forma convergente, a aquisição do plasmídeo virulento pYV/pCD1, que codifica o sistema de secreção do tipo III (T3SS) Ysc-Yop, essencial para a virulência desses patógenos (Bottone, 1999).

Ainda, outras espécies desempenham papel importante como patógenos em hospedeiros não mamíferos, como *Y. ruckeri* (Wrobel *et al.*, 2019), associada a infecções sistêmicas em peixes, e *Y. entomophaga*, patógeno de insetos (Hurst *et al.*, 2011). Nessas espécies, a patogenicidade está associada a diferentes fatores de virulência adaptados aos seus respectivos hospedeiros. Em *Y. entomophaga*, destaca-se o complexo tóxico Yen-TC, um sistema multiproteico codificado em uma ilha de patogenicidade cromossômica descrito como um dos principais

determinantes da atividade entomopatogênica, com ação em larvas de *Costelytra zealandica* (Scarabaeidae: Melolonthinae) (Hurst et al., 2011).

A partir desses exemplos, infere-se que a patogenicidade no gênero *Yersinia* não está restrita a um único grupo evolutivo, bem como não depende de um conjunto conservado de determinantes genéticos, o que reflete diferentes trajetórias de adaptação a hospedeiros e nichos ecológicos (Reuter et al., 2014; Alenizi et al., 2016; Hurst et al., 2011; Wrobel et al., 2019). No entanto, apesar do amplo conhecimento acerca das espécies de importância médica, a diversidade do gênero *Yersinia* permanece pouco caracterizada. Espécies ambientais ou associadas a hospedeiros não humanos ainda são subamostradas e frequentemente negligenciadas em estudos ecológicos e genômicos, resultando em lacunas quanto a sua distribuição e diversidade genética. Apesar disso, a ampla diversidade de hospedeiros e nichos ecológicos torna o gênero um importante modelo para investigar processos de diversificação e adaptação bacteriana.

1.2 Plasticidade genômica e evolução bacteriana

Em bactérias, processos adaptativos geralmente envolvem alterações no repertório gênico das linhagens, o que contribui para a diversidade observada nesses microrganismos e pode influenciar características relacionadas à interação com hospedeiros, utilização de recursos ambientais e ocupação de diferentes nichos ecológicos (Arnold; Huang; Hanage, 2022; McInerney; McNally; O'Connell, 2017). Nas últimas duas décadas, a crescente disponibilidade de genomas bacterianos, decorrente dos avanços da genômica, impulsionou o desenvolvimento da genômica comparativa e permitiu a investigação dos padrões evolutivos relacionados à variação no seu conteúdo gênico (McInerney; McNally; O'Connell, 2017). A partir desses trabalhos, constataram-se as limitações, em termos de diversidade genética, do estudo de um único genoma, tendo em vista que mesmo indivíduos intimamente relacionados filogeneticamente diferem quanto ao repertório gênico (Tettelin et al., 2005).

Nesse contexto, surgiu o conceito de “pangenoma”, definido como o conjunto de genes e variantes genômicas encontradas em múltiplos

representantes de uma espécie, de modo a propiciar uma compreensão mais abrangente da diversidade genética microbiana de um dado conjunto (Golicz *et al.*, 2020; Matthews *et al.*, 2024). Embora originalmente proposto para espécies, esse conceito tem sido expandido para análises em subconjuntos ou em escalas taxonômicas mais amplas, como o gênero. Essa coleção de genes pode ser subdividida em genes *core*— compartilhados por todos os genomas, e genes acessórios— presentes somente em parte dos genomas amostrados (Golicz *et al.*, 2020). É neste último compartimento que se concentra a variabilidade genética do grupo estudado (Tettelin *et al.*, 2005).

A dinâmica do genoma acessório é diretamente influenciada pela transferência horizontal de genes (*Horizontal Gene Transfer*, HGT), um dos principais mecanismos promotores de fluxo gênico em procariotos (Arnold; Huang; Hanage, 2022). Através desse processo, diferentes linhagens compartilham genes por meio de elementos genéticos móveis (*Mobile Genetic Elements*, MGEs), incluindo plasmídeos, transposons, ilhas genômicas e bacteriófagos, o que possibilita uma rápida e ampla diversificação do repertório gênico desses microrganismos (Brockhurst *et al.*, 2019). Genes associados à especialização metabólica, resistência a antimicrobianos, interação com hospedeiros e fatores de virulência são comumente mobilizados, influenciando a adaptação dos recipientes a novos ambientes e nichos ecológicos (McInerney; McNally; O'Connell, 2017).

Além da HGT mediada por MGEs, a recombinação homóloga— processo no qual fragmentos de DNA são incorporados ao genoma de modo a substituir sequências homólogas pré-existentes, formando genomas mosaico—, representa um importante mecanismo de diversificação genômica em bactérias (Michod *et al.*, 2008; Didelot; Maiden, 2010). Como resultado, esse processo contribui para a diversidade genética entre as linhagens e pode alterar funções celulares e fenótipos associados à interação com os hospedeiros ou ao uso de recursos ambientais (Didelot; Maiden, 2010).

Adicionalmente, processos de perda gênica também são importantes mecanismos de adaptação bacteriana (Akeno *et al.*, 2014). Em linhagens associadas a nichos mais especializados, como organismos intracelulares e simbioses, a redução gênica progressiva pode decorrer do relaxamento seletivo e

da deriva genética, resultando na perda de funções não essenciais àquele contexto ecológico e em profunda especialização (Gil *et al.*, 2002). Esse fenômeno é observado, por exemplo, em bactérias intracelulares obrigatórias do gênero *Rickettsia*, cujos genomas apresentam tamanho reduzido e elevados índices de pseudogenização, bem como perda de vias metabólicas supridas pela célula hospedeira (Diop; Raoult; Fournier, 2018).

Semelhantemente, em *Y. pestis*, processos de pseudogenização estiveram associados à transição para um novo estilo de vida patogênico (Parkhill *et al.*, 2001). Demeure e colaboradores demonstraram que essa espécie emergiu recentemente de um clone de *Y. pseudotuberculosis* há cerca de 6.000 anos, pouco antes das primeiras pandemias de peste registradas em humanos (Demeure *et al.*, 2019; Carcauzon *et al.*, 2019). Tal episódio foi acompanhado pela perda ou inativação de genes metabólicos e envolvidos na sobrevivência intestinal e, conseqüentemente, intensa redução genômica (Demeure *et al.*, 2019; Achtman *et al.*, 1999). Em paralelo, a espécie passou por um processo de adaptação a um novo nicho ecológico, marcado pela transmissão vetorial e pela infecção sistêmica no hospedeiro vertebrado, o que foi possibilitado pela aquisição horizontal de elementos genéticos móveis, associados à colonização no vetor e à disseminação no hospedeiro (Chen *et al.*, 2014). Tais exemplos demonstram que a plasticidade genômica bacteriana não se limita à expansão do repertório gênico, mas envolve também processos de perda e reorganização funcional associados à adaptação a diferentes nichos ecológicos e estilos de vida.

1.3 Diversidade ecológica e adaptação bacteriana

A diversidade genética observada nos procariotos constitui simultaneamente um produto do processo evolutivo e um determinante da dinâmica ecológica dessas populações (Brockhurst *et al.*, 2019). Na natureza, esses microrganismos participam de redes ecológicas complexas— envolvendo hospedeiros, competidores, predadores e disponibilidade diferencial de recursos ambientais—, cujas interações são diretamente moduladas pela variação genética (Xu, 2006).

Em contrapartida, a dinâmica do fluxo gênico bacteriano também parece ser influenciada pelo contexto ecológico em que essas populações estão inseridas. Cordero e Polz argumentam que a transferência horizontal de genes tende a ocorrer com maior frequência entre linhagens que compartilham nichos ecológicos semelhantes ou apresentam oportunidades recorrentes de contato, de modo que a conectividade ecológica entre populações estrutura a circulação de genes no pangenoma (Cordero; Polz, 2014).

Assim, processos como competição, cooperação e predação por bacteriófagos podem favorecer diferentes variantes gênicas em escalas espaciais reduzidas, impedindo a fixação uniforme de determinados genótipos (Dalziel; Rogers; Schulte, 2009). Como muitos desses fenótipos ecológicos estão associados ao genoma acessório, a dinâmica dessas interações influencia diretamente a frequência e disseminação de diferentes *loci* nas populações bacterianas (Cordero; Polz, 2014). Desse modo, enquanto alguns genes acessórios permanecem amplamente distribuídos em múltiplas linhagens, outros apresentam distribuição restrita ou transitória, refletindo contrastes na intensidade e estabilidade das pressões seletivas atuantes em cada contexto ecológico (Dewar *et al.*, 2024). Isso ocorre porque a aquisição de novos genes não necessariamente resulta em vantagem adaptativa, uma vez que a manutenção desses elementos depende da compatibilidade funcional com o *background* genético do organismo receptor e dos possíveis custos associados à incorporação de DNA exógeno (Wiedenbeck; Cohan, 2011).

Em escalas evolutivas mais amplas, essas dinâmicas ecológicas também se refletem na própria organização do pangenoma bacteriano. Dewar *et al.* demonstraram que bactérias associadas a ambientes ecologicamente heterogêneos ou submetidas a múltiplas interações bióticas tendem a apresentar repertórios gênicos mais amplos e dinâmicos, enquanto linhagens altamente especializadas frequentemente exibem conjuntos gênicos mais restritos e estáveis (Dewar *et al.*, 2024). Dessa forma, fatores ecológicos influenciam desde a dinâmica de genes individuais até a organização evolutiva do pangenoma bacteriano.

Em *Yersinia*, essas dinâmicas são evidenciadas pela coexistência de espécies ecologicamente sobrepostas, ainda que pertencentes a linhagens evolutivas distintas. *Y. enterocolitica* e *Y. pseudotuberculosis*, por exemplo, podem ser isoladas dos mesmos reservatórios animais, uma vez que estão associadas ao trato gastrointestinal de vertebrados, compartilhando pressões ecológicas relacionadas à colonização intestinal e à interação com hospedeiros (Platt-Samoraji *et al.*, 2019; Galindo *et al.*, 2011). Além disso, ambas são consideradas patógenos transmitidos por alimentos, conectando animais, ambiente, cadeias alimentares e infecção humana (Fukushima; Schimizu; Inatsu, 2011).

Entretanto, apesar da sobreposição ecológica e das oportunidades recorrentes de contato entre diferentes linhagens de *Yersinia*, a distribuição de fatores associados à virulência, por exemplo, permanece altamente heterogênea. Isso se evidencia dentro da própria espécie *Y. enterocolitica*, na qual diferentes biotipos podem ocorrer em contextos ecológicos associados a hospedeiros, incluindo isolados portadores ou não do plasmídeo de virulência pYV, embora apresentem diferenças marcantes no repertório gênico acessório (Reuter *et al.*, 2014; Alenizi *et al.*, 2016). De forma semelhante, embora espécies filogeneticamente próximas associadas a ambientes não clínicos, como *Y. intermedia*, *Y. frederiksenii* e *Y. kristensenii*, apresentem diversidade genômica significativa, sua relevância funcional e seu papel na patogenicidade humana permanecem não totalmente esclarecidos (Nguyen *et al.*, 2020; Fukushima; Schimizu; Inatsu, 2011).

Em conjunto, esses padrões sugerem que a diversidade ecológica observada no gênero *Yersinia* está associada a diferentes trajetórias evolutivas do conteúdo gênico, nas quais linhagens sob pressões ecológicas e seletivas distintas podem apresentar variações no repertório gênico acessório, refletindo diferenças em capacidade de colonização de hospedeiros, persistência ambiental e potencial patogênico. Entretanto, ainda são limitados os estudos comparativos que integrem a estrutura filogenômica de *Yersinia* com a dinâmica do pangenoma e com a diversidade de nichos ecológicos observados no gênero.

2. HIPÓTESE

A organização do pangenoma de *Yersinia* reflete a estrutura filogenética das linhagens e a heterogeneidade ecológica observada entre os isolados do gênero.

3. OBJETIVOS

3.1 Objetivo Geral

Investigar a organização do pangenoma do gênero *Yersinia* e sua relação com a estrutura filogenética e o contexto ecológico dos isolados.

3.2 Objetivos Específicos

- Inferir a estrutura filogenômica e a delimitação taxonômica do gênero *Yersinia*;
- Caracterizar a estrutura do pangenoma do gênero *Yersinia* e analisar a anotação funcional do genoma persistente e dos genes exclusivos de suas respectivas espécies;
- Associar os dados genômicos com as categorias ecológicas associadas aos isolados.

4. METODOLOGIA

4.1 Obtenção e curadoria dos genomas

Para análise comparativa, todos os genomas disponíveis de *Yersinia* spp. foram recuperados do banco de dados do National Center for Biotechnology Information (NCBI) em 29/03/2026, utilizando a ferramenta de linha de comando NCBI Datasets (Sayers *et al.*, 2025). Inicialmente, foram excluídos genomas montados a partir de metagenomas (*Metagenome-Assembled Genomes*, MAGs) e registros considerados atípicos ou inconsistentes com os critérios de inclusão estabelecidos, resultando em um conjunto de 9.955 genomas. Para reduzir a redundância genômica e garantir a diversidade do conjunto de dados, os genomas foram deduplicados utilizando o *software* skani (Shaw; Yu, 2023), com base em critérios de similaridade genômica. Foram considerados redundantes os genomas que apresentaram identidade média de nucleotídeos (*Average Nucleotide Identity*, ANI) $\geq 99,5\%$ e fração alinhada (*Aligned Fraction*, AF) $\geq 95\%$ (Rodriguez *et al.*, 2023), sendo agrupados em *clusters* de alta similaridade. Para cada *cluster*, foi selecionado um genoma representativo com base em critérios de qualidade e confiabilidade da montagem. As métricas de completude e contaminação, bem como estatísticas de qualidade, incluindo número de *contigs* e N50, foram obtidas a partir do CheckM2 v1.1.0 (Chklovski *et al.*, 2023) e utilizadas como critérios de seleção. Genomas classificados como estirpes-tipo foram priorizados sempre que disponíveis. Adicionalmente, foram excluídos genomas com contaminação $\geq 5\%$ e completude $\leq 90\%$, resultando em um conjunto final de 1.013 genomas para as análises subsequentes.

4.2 Reconstrução filogenética e delimitação taxonômica do gênero *Yersinia*

A reconstrução filogenética foi realizada por máxima verossimilhança a partir de um alinhamento concatenado de genes ortólogos bacterianos. Os genes ortólogos foram identificados utilizando o *pipeline* UBCG2 (Kim *et al.*, 2021), o qual seleciona um conjunto de genes marcadores bacterianos amplamente conservados para análises filogenômicas. Cada ortólogo foi alinhado individualmente utilizando MAFFT v7.505 (Katoh & Standley, 2013), por meio da implementação integrada no *pipeline* UBCG2. Em seguida, os alinhamentos

individuais foram concatenados para geração de uma matriz *multilocus*, utilizada na inferência filogenética. A árvore foi reconstruída utilizando IQ-TREE v3.0.1 (Minh et al., 2020), com seleção automática do melhor modelo evolutivo pelo ModelFinder (Kalyvaanamoorthy et al., 2017). O suporte dos ramos foi estimado com 1.000 replicações de *ultrafast bootstrap* (UFBoot) e 1.000 testes *SH-like approximate likelihood ratio test* (SH-aLRT). A árvore foi enraizada utilizando as estirpes-tipo de *Serratia marcescens* (GCF_004684145.1) e *Ewingella americana* (GCF_008693045.1) como grupos externos (*outgroups*), representando, respectivamente, o grupo-irmão de *Yersinia* e o grupo-irmão do clado formado por *Yersinia* e *S. marcescens*.

Para avaliar a delimitação de espécies dentro do gênero *Yersinia*, foi realizada uma análise de identidade média de nucleotídeos (Average Nucleotide Identity, ANI) entre os 1.013 genomas não redundantes. As comparações par a par foram utilizadas para construir uma matriz de ANI, a partir da qual os genomas foram agrupados por relações de ANI $\geq 95\%$ (Jain et al., 2018). Cada *cluster* de ANI representa uma espécie genômica proposta. Os agrupamentos obtidos foram comparados às designações taxonômicas disponíveis nos registros do NCBI, permitindo identificar concordâncias e genomas potencialmente classificados de forma inadequada e reclassificá-los.

4.3 Construção e estrutura do pangenoma

Para avaliar a estrutura genômica no nível do conteúdo gênico, foi realizada uma análise exploratória do pangenoma utilizando o PPanGGOLiN v2.3.0 (Gautreau et al., 2020), a partir de genomas de *Yersinia* spp. em formato FASTA (.fna), previamente obtidos e filtrados na etapa de desreplicação. A predição e anotação gênica foram realizadas internamente pela ferramenta utilizando Pyrodigal (Larralde, 2022). As famílias de genes ortólogos foram definidas com base em 80% de identidade e cobertura de sequência. A partir dessas informações, foi construído um grafo de pangenoma, no qual cada nó representa uma família gênica e as arestas indicam a adjacência entre genes nos genomas.

O particionamento do pangenoma foi realizado pelo próprio PPanGGOLiN em três categorias: *persistent* (correspondente ao genoma *core* na nomenclatura adotada pela ferramenta), *shell* e *cloud*, com base na frequência de ocorrência

das famílias gênicas entre os genomas. As famílias *persistent* correspondem a genes altamente conservados, enquanto *shell* e *cloud* representam famílias de frequência intermediária e rara, respectivamente.

A matriz de presença/ausência gênica foi utilizada para análises estatísticas no ambiente R v4.4.3. A estrutura genômica foi explorada por meio de análise de coordenadas principais (PCoA), baseada em distâncias de Jaccard (Prokopenko et al., 2016). As diferenças na composição gênica entre os *clusters* foram avaliadas por meio de uma análise de variância permutacional multivariada (PERMANOVA). Como a significância da PERMANOVA pode ser influenciada por diferenças na dispersão dos grupos, a homogeneidade das dispersões multivariadas foi testada utilizando uma análise permutacional de dispersões multivariadas (PERMDISP), baseada nas distâncias genômicas aos respectivos centróides. Quando detectadas diferenças significativas, comparações pareadas entre os grupos foram realizadas utilizando teste de TukeyHSD (*Honestly Significant Difference*) aplicado às distâncias do centróide. Assim, a interpretação dos resultados da PERMANOVA foi realizada em conjunto com o PERMDISP.

Em seguida, para estimar as curvas de acumulação do pangenoma, foram geradas 100 permutações aleatórias da ordem de inclusão dos genomas. Para cada ordenamento, foram calculados o número total de genes únicos (tamanho do pangenoma) e o número de genes presentes em todos os genomas incluídos (tamanho do genoma *core*). As curvas finais de acumulação representam a média entre as permutações, corrigindo o viés associado à ordem de amostragem. A abertura do pangenoma foi avaliada por meio do ajuste da lei de Heap à curva média de acumulação do pangenoma e seu desvio padrão, conforme descrito por Tettelin e colaboradores (Tettelin et al., 2008). O modelo de lei de potência relaciona o número de genomas analisados (N) ao tamanho do pangenoma ($P(N)$) segundo a expressão: $P(N) = kN^\gamma$. Os parâmetros do modelo foram estimados por regressão linear em escala log-log, e a abertura do pangenoma foi inferida a partir do expoente $\alpha = 1 - \gamma$ em que $\alpha < 1$ indica um pangenoma aberto.

4.4 Anotação funcional

Para a análise funcional, foram consideradas duas abordagens complementares. Em um primeiro nível, o genoma persistente do conjunto

completo de genomas de *Yersinia* spp. foi analisado de forma agregada, representando o repertório gênico conservado do gênero. Em um segundo nível, foram recuperados os genes exclusivos de cada espécie (*cluster* ANI), sendo o complexo *Y. pseudotuberculosis* analisado separadamente, uma vez que seus membros são tratados como espécies distintas na literatura. A anotação funcional foi realizada utilizando o eggNOG-mapper v2.0.8 na plataforma Galaxy, com o algoritmo de alinhamento DIAMOND v2.0.7 e parâmetros padrão. Os genes persistentes anotados foram mapeados para categorias de Grupos de Ortólogos (COGs), sendo que genes associados a múltiplos COGs foram contabilizados em todas as categorias correspondentes. Já para os genes exclusivos foi realizada uma curadoria manual quanto à associação com elementos genéticos móveis baseada na anotação funcional. Quando aplicável, os genes foram classificados em: transposases e sequências de inserção (IS), elementos integrativos, proteínas associadas a bacteriófagos (estruturais e de montagem), genes relacionados à lise fágica, componentes plasmidiais, sistemas de conjugação e ilhas genômicas.

4.5 Classificação ecológica dos isolados

Para caracterizar os contextos de procedência dos genomas de *Yersinia* spp., foram obtidos os metadados disponíveis nos registros do NCBI *BioSample* por meio dos respectivos números de acesso. Foram coletadas, quando disponíveis, as informações referentes aos campos *host* e *isolation source*, os quais foram compilados e tratados de forma independente. As descrições presentes nesses campos foram revisadas manualmente e organizadas em uma classificação hierárquica composta por três níveis de detalhamento, baseada nas informações registradas nos metadados originais.

No primeiro nível (origem do isolado), os isolados foram agrupados em categorias amplas de procedência, incluindo hospedeiro, ambiente, alimento e clínico. No segundo nível (compartimento ecológico), as categorias foram refinadas com base em descrições mais específicas, derivadas do campo *isolation source*, incluindo, ambiente aquático, esgoto, solo, alimentos de origem animal e derivados, alimentos de origem vegetal e derivados, ambiente intestinal, sangue e tecido. No terceiro nível (hospedeiro), as informações provenientes do campo *host* foram utilizadas para identificar os grupos de hospedeiros associados aos isolados, incluindo humanos, demais mamíferos, peixes, aves e insetos. Registros

sem informação ou cuja descrição foi insuficiente para permitir a classificação em determinado nível foram categorizados como "não determinado", conforme aplicável.

Os metadados classificados foram integrados às análises de estrutura de ordenação do pangenoma. Essa integração permitiu avaliar a distribuição dos diferentes contextos ecológicos no espaço multivariado definido pelo conteúdo gênico dos genomas de *Yersinia spp.* Adicionalmente, a origem de isolamento dos genomas foi mapeada sobre a árvore filogenômica do gênero. Essas análises foram utilizadas de forma exploratória para investigar a relação entre a estrutura do pangenoma, o contexto ecológico dos isolados e a organização filogenética do gênero *Yersinia*.

5. RESULTADOS

5.1 Impacto da desreplicação na representatividade genômica de *Yersinia*

Foram recuperados 9.955 genomas pertencentes ao gênero *Yersinia* disponíveis no NCBI. A distribuição entre as espécies foi assimétrica, com predominância de *Y. enterocolitica* (5.375 genomas) e *Y. pestis* (3.638 genomas), que em conjunto representaram aproximadamente 90,5% dos acessos. As demais espécies apresentaram baixa representatividade, variando de um a algumas centenas de genomas.

Após a desreplicação, foram obtidos 1.013 genomas não redundantes, correspondendo a uma redução de quase 90% em relação ao conjunto inicial. Esse padrão variou entre as espécies, sendo particularmente pronunciado em *Y. pestis*, cujo número de genomas foi reduzido para apenas 11 (0,3% dos registros originais). Em contraste, *Y. enterocolitica* manteve-se como a espécie mais representativa após a deduplicação, com 688 genomas (68% do conjunto final). As demais espécies apresentaram reduções intermediárias, compatíveis com seus níveis de amostragem e diversidade intraespecífica.

Espécie	Nº genomas depositados no NCBI	Nº genomas após deduplicação	Percentual de retenção (%)
<i>Y. aldovae</i>	11	10	90,91
<i>Y. aleksiciae</i>	15	11	73,33
<i>Y. alsatica</i>	20	10	50
<i>Y. artesisana</i>	4	1	25
<i>Y. bercovieri</i>	21	16	76,19
<i>Y. canariae</i>	3	3	100
<i>Y. enterocolitica</i>	5.375	688	12,80
<i>Y. entomophaga</i>	3	2	66,67
<i>Y. frederiksenii</i>	25	19	76
<i>Y. fenwicki</i>	1	1	100
<i>Y. hibernica</i>	4	4	100

<i>Y. intermedia</i>	48	36	75
<i>Y. kristensenii</i>	55	26	47,27
<i>Y. massiliensis</i>	17	14	82,35
<i>Y. mollaretii</i>	37	14	37,84
<i>Y. nurmii</i>	2	1	50
<i>Y. pekkanenii</i>	3	1	33,33
<i>Y. pestis</i>	3.638	11	0,3
<i>Y. proxima</i>	19	7	36,84
<i>Y. pseudotuberculosis</i>	370	66	17,84
<i>Y. rochesterensis</i>	14	6	42,86
<i>Y. rohdei</i>	13	12	92,31
<i>Y. ruckeri</i>	208	25	12,02
<i>Y. similis</i>	10	8	80
<i>Y. thracica</i>	4	2	50
<i>Y. vastinensis</i>	10	4	40
<i>Y. wautersii</i>	3	2	66,67
<i>Yersinia sp.</i>	22	18	81,82

Quadro 1: Impacto da desreplcação genômica na representatividade das espécies de *Yersinia*.

5.2 Estrutura taxonômica e filogenômica do gênero *Yersinia*

A análise do conjunto desreplcado utilizando o limiar de ANI $\geq 95\%$ permitiu a identificação de agrupamentos genômicos compatíveis com os limites de espécie geralmente adotados para procariotos com base na similaridade nucleotídica média. A matriz de ANI revelou estruturação consistente entre os agrupamentos, com elevada similaridade nucleotídica intra-*cluster* (média de 98,55%) e menor identidade entre *clusters* distintos (média de 85,98%) (Figura 1).

A partir disso, foram definidos 25 *clusters* genômicos, cuja identidade taxonômica foi atribuída com base na presença de estirpes-tipo, resultando na designação de *Y. artesianae*, *Y. aldovae*, *Y. aleksiciae*, *Y. alsatica*, *Y. canariae*, *Y.*

enterocolitica, *Y. entomophaga*, *Y. intermedia*, *Y. frederiksenii*, *Y. fenwickii*, *Y. hibernica*, *Y. kristensenii*, *Y. massiliensis*, *Y. mollaretii*, *Y. nurmii*, *Y. pekkanenii*, *Y. proxima*, *Y. pseudotuberculosis* (incluindo *Y. pestis*, *Y. similis* e *Y. wautersii*), *Y. rochesterensis*, *Y. rohdei*, *Y. ruckeri*, *Y. thracica* e *Y. vastinensis*. Genomas com classificação taxonômica inconsistente foram reavaliados e reatribuídos com base na sua associação ao respectivo *cluster* contendo a estirpe-tipo (Quadro A1).

Notavelmente, *Y. pseudotuberculosis*, *Y. pestis*, *Y. wautersii* e *Y. similis* foram recuperadas em um mesmo *cluster* definido pelo critério de ANI $\geq 95\%$ e formaram um único clado na reconstrução filogenética. Dessa forma, esses táxons foram tratados como pertencentes a um mesmo agrupamento genômico (complexo *Y. pseudotuberculosis*) no contexto deste estudo.

Adicionalmente, dos 18 genomas previamente anotados como *Yersinia* sp., 7 foram redistribuídos para *clusters* correspondentes a espécies já descritas com base na similaridade genômica, substituindo a classificação original do NCBI (Quadro A1). Os 11 genomas remanescentes agruparam-se em um único *cluster* independente na análise de ANI. Esse agrupamento também foi recuperado como um clado fortemente sustentado na reconstrução filogenética (bootstrap = 100). Assim, enquanto os demais *clusters* puderam ser associados a espécies reconhecidas do gênero *Yersinia* por meio de suas respectivas estirpes-tipo, esse grupo permaneceu sem correspondência com qualquer estirpe-tipo incluída na análise, sugerindo a existência de uma linhagem genômica distinta e ainda não descrita dentro do gênero analisado.

Adiante, a reconstrução filogenética dos 1.013 genomas recuperou uma topologia congruente com os agrupamentos definidos pela análise de ANI (Figura 2). Dos 25 *clusters* delimitados por ANI, 24 foram recuperados como grupos monofiléticos na árvore filogenética. A exceção correspondeu a *Y. fenwickii*, cujo genoma foi inserido no interior do clado de *Y. enterocolitica*.

Além disso, a projeção dos dados ecológicos de origem de isolamento na árvore evidenciou padrões de distribuição distintos entre os táxons. *Y. enterocolitica* apresentou distribuição ubíqua, com isolados recuperados de amostras clínicas, alimentares e ambientais, além de uma diversidade de hospedeiros mamíferos (incluindo humanos), sugerindo elevada heterogeneidade ecológica. Por outro lado, o *cluster* correspondente a *Yersinia* sp. esteve predominantemente associado a isolados oriundos de hospedeiros mamíferos

considerando o conjunto analisado. De forma similar, *Y. ruckeri* teve todos os isolados analisados provenientes de peixes. Já, em relação às características genômicas, observou-se variação no tamanho dos genomas entre os *clusters*, com destaque para *Y. ruckeri* e *Y. aldovae*, que apresentaram genomas relativamente maiores em comparação aos demais grupos.

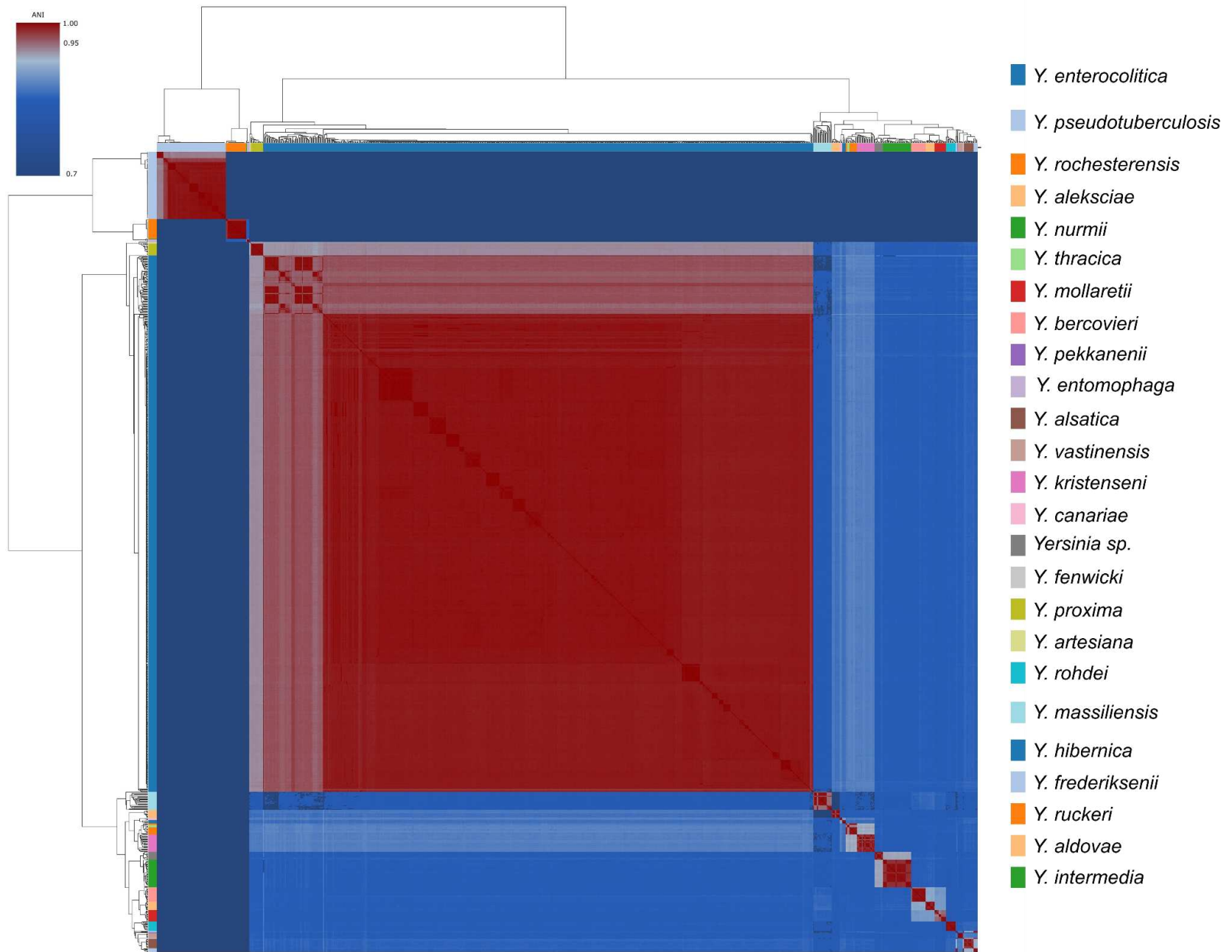


Figura 1: *Heatmap* de identidade nucleotídica média (ANI) entre genomas de *Yersinia*. Cada célula representa a ANI entre pares de genomas. Os agrupamentos genômicos foram delimitados utilizando o limiar de ANI $\geq 95\%$, resultando em 25 *clusters*. Os dendrogramas representam o agrupamento hierárquico aplicado à matriz de similaridade para fins de visualização. Observa-se elevada similaridade intra-*cluster* (ANI média = 98,55%) e menor identidade entre *clusters* distintos (ANI média = 85,98%).

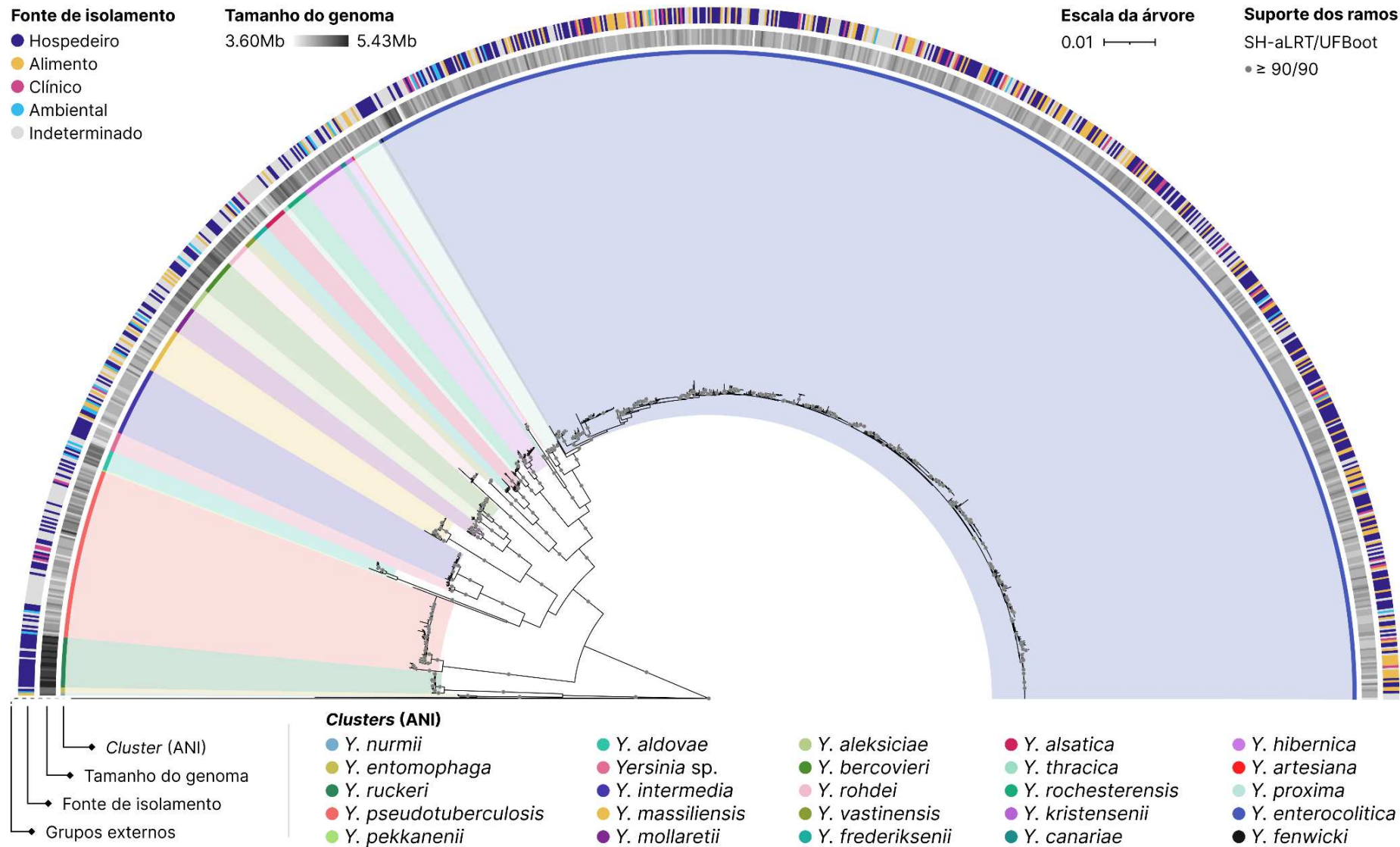


Figura 2: Estrutura filogenômica do gênero *Yersinia*. Árvore de máxima verossimilhança reconstruída a partir de 1.013 genomas de *Yersinia*, utilizando o IQ-TREE, com *S. marcescens* e *E. americana* como grupos externos. Os valores de suporte correspondem às métricas SH-aLRT ≥ 90 e UFBoot ≥ 90 . As anotações, do interior para o exterior da árvore, correspondem a: (1) *clusters* genômicos delimitados por ANI ≥ 95 ; (2) tamanho do genoma; e (3) fonte de isolamento. Os setores coloridos delimitam os agrupamentos genômicos identificados pelas análises de ANI.

5.3 Estrutura, diversidade e dinâmica do pangenoma de *Yersinia*

O pangenoma de *Yersinia* apresentou uma distribuição assimétrica na frequência gênica. Com um total de 58.301 genes, o pangenoma foi segmentado em 2.458 genes persistentes (4,22%), 10.298 genes *shell* (17,66%) e 45.545 genes *cloud* (78,12%) (Quadro 2). Os genes acessórios (*shell* e *cloud*), em conjunto, representaram quase a totalidade do pangenoma (95,78%), destacando a predominância do conteúdo genômico acessório. A análise de acumulação do pangenoma revelou uma taxa de crescimento elevada no tamanho do pangenoma, enquanto o número de genes persistentes rapidamente atinge um *plateau* (Figura 3). Esse padrão sugere a estabilização do genoma persistente, juntamente a um conjunto altamente dinâmico de genes acessórios que continua a se expandir com o aumento da amostragem. O ajuste da lei de Heaps à curva de acumulação resultou em um valor α de $0,621 \pm 0,012$ ($\gamma = 0,379 \pm 0,012$), evidenciando um pangenoma aberto, no qual a incorporação de novos genomas tende a adicionar continuamente novas famílias gênicas.

Partição	Número de famílias gênicas	Percentual (%)
<i>Persistent</i>	2458	4.22
<i>Shell</i>	10298	17.66
<i>Cloud</i>	45545	78.12
Total	58301	100

Quadro 2: Distribuição das famílias gênicas nas partições do pangenoma de *Yersinia*

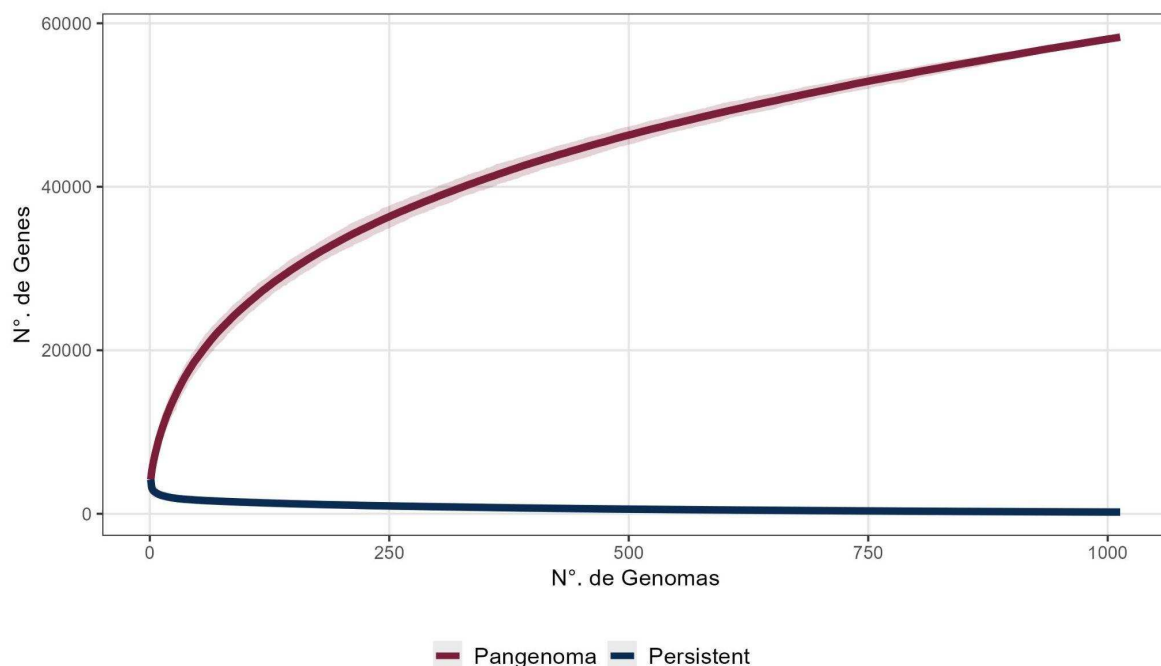


Figura 3. Dinâmica de crescimento do pangenoma de *Yersinia*. Curvas de acumulação do pangenoma e do genoma persistente obtidas a partir de permutações aleatórias da ordem de adição dos genomas, ajustadas pela lei de Heaps. As áreas sombreadas representam a média $\pm$ desvio padrão.

A análise de coordenadas principais (PCoA) baseada na matriz de presença e ausência de genes revelou uma estruturação consistente do conteúdo gênico entre os genomas de *Yersinia* (Figura 4). Os dois primeiros eixos da ordenação explicaram conjuntamente 37,5% da variação observada no conteúdo gênico (Coord1 = 27,5%; Coord2 = 10,0%). A ordenação evidenciou a formação de três *clusters* do pangenoma —assim denominados para distingui-los dos clusters definidos por ANI— no espaço do repertório gênico, cuja separação foi suportada pela análise de PERMANOVA (Figura 4, $R^2 = 0,343$; $F = 264,03$; $p < 0,001$). Ainda, a análise de PERMDISP indicou diferenças significativas na dispersão dos grupos. Comparações pareadas das distâncias ao centróide mostraram que o *Cluster* 1 apresentou menor dispersão em relação ao *Cluster* 3 (Figura 4, p ajustado = 0,00016), enquanto o *Cluster* 2 apresentou dispersão intermediária e não diferiu significativamente dos demais *clusters* do pangenoma.

O *Cluster* 1 (do pangenoma), composto por *Y. pestis*, *Y. pseudotuberculosis*, *Y. similis* e *Y. wautersii*, corresponde a um complexo genômico também suportado por análises de ANI e reconstrução filogenética, caracterizando um grupo evolutivo consistente em diferentes níveis de

organização genômica. De modo semelhante, o *Cluster 2* (do pangenoma), formado por *Y. ruckeri*, *Y. entomophaga* e *Y. nurmii*, corresponde a um agrupamento monofilético na filogenia, indicando concordância entre conteúdo gênico e história evolutiva. Por outro lado, o *Cluster 3* (do pangenoma) agrupou as demais espécies do gênero e apresentou maior dispersão no espaço multivariado, refletindo maior heterogeneidade no repertório gênico e representando um conjunto mais diverso de linhagens dentro de *Yersinia*. Dentro desse conjunto, *Y. enterocolitica* apresentou posicionamento mais extremo ao longo do primeiro eixo da ordenação, resultando em uma maior divergência no espaço de conteúdo gênico em relação às demais espécies do *cluster*.

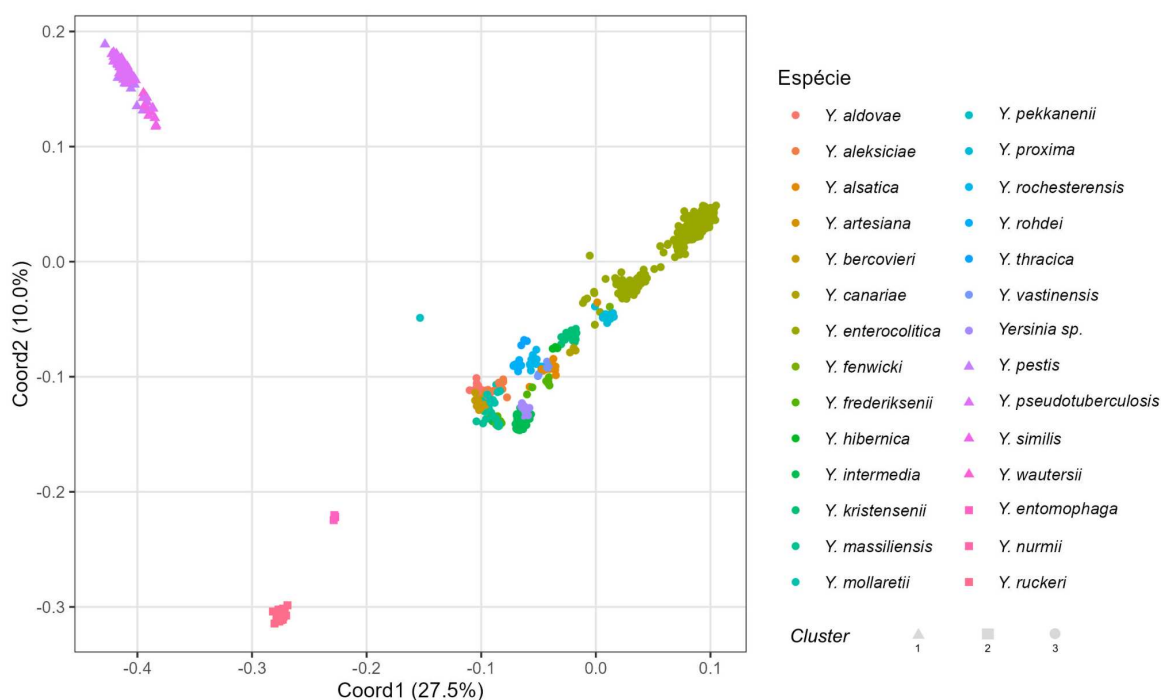


Figura 4. Diversidade do conteúdo genético em *Yersinia* inferida a partir da análise do pangenoma. A Análise de Coordenadas Principais (PCoA), baseada em dados de presença/ausência de genes, evidencia a separação dos genomas em três *clusters* distintos (PERMANOVA, $R^2 = 0,343$; $F = 264,03$; $p < 0,001$), considerando 1.013 genomas. As cores representam as diferentes espécies e as formas indicam os *clusters* genômicos encontrados.

5.4 Anotação funcional do genoma persistente e dos genes exclusivos das espécies de *Yersinia*

Em seguida, foi analisada a composição funcional do genoma persistente de *Yersinia*. Os ortólogos pertencentes ao genoma persistente foram classificados em diferentes categorias funcionais COG (Figura 5). Embora a categoria COG S

(função desconhecida) tenha sido a mais abundante individualmente, as categorias relacionadas ao metabolismo representaram, em conjunto, a maior fração do genoma persistente, totalizando 1.131 genes (46%). Além das funções metabólicas, o genoma persistente apresentou elevada representatividade de genes associados ao armazenamento e processamento de informações, incluindo transcrição (COG K; 182 genes), tradução, estrutura ribossomal e biogênese (COG J; 192 genes), bem como replicação, recombinação e reparo (COG L; 95 genes). Também foram identificados genes envolvidos na biogênese da parede celular e da membrana (COG M; 164 genes).

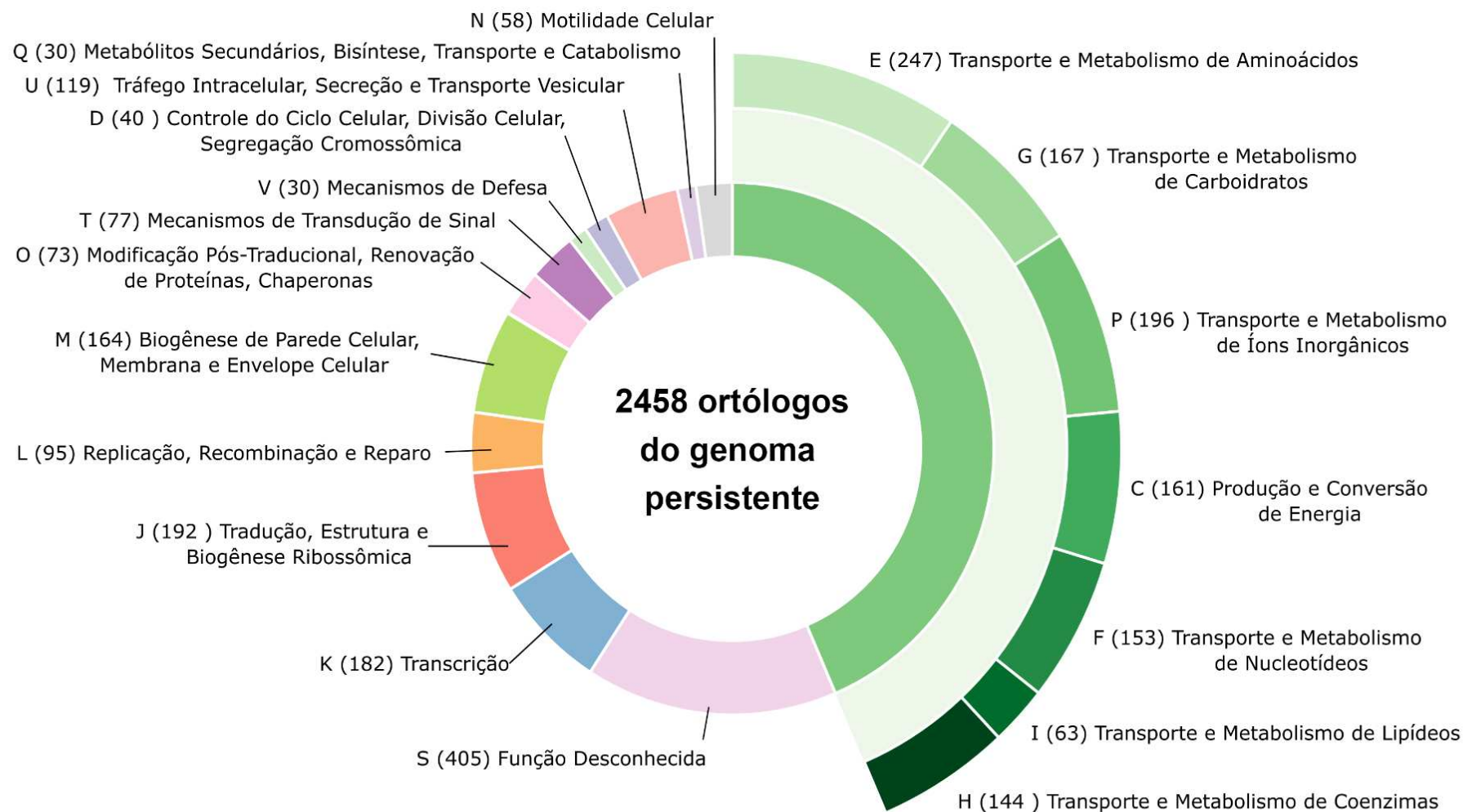


Figura 5. Distribuição funcional dos ortólogos do genoma persistente do gênero *Yersinia*. Visão circular dos 2.458 ortólogos compartilhados do genoma persistente, classificados de acordo com as categorias funcionais do COG. O círculo interno indica o número total de genes por categoria, enquanto o anel externo destaca as subcategorias metabólicas.

Quanto ao genoma acessório, os genes exclusivos das espécies de *Yersinia* foram anotados funcionalmente e classificados, quando cabível, quanto à associação com MGEs (Figura 6). O número total de genes exclusivos variou substancialmente entre os grupos, desde algumas centenas a dezenas de milhares de genes, refletindo diferenças no número de genomas representados em cada grupo, bem como a diversidade intraespecífica. Não foram identificados genes exclusivos para *Y. wautersii* no conjunto analisado. Em relação à composição funcional, os MGEs corresponderam a aproximadamente 6–18% do conteúdo exclusivo entre as espécies.

Entre os genes classificados como MGEs, componentes estruturais de fago representaram a categoria mais abundante na maioria das espécies analisadas. Transposons e sequências de inserção (IS), assim como elementos integrativos e conjugativos, componentes plasmidiais e elementos associados à lise fágica também foram identificados, embora em proporções variáveis entre as espécies. Por outro lado, ilhas genômica foram identificadas apenas em algumas espécies, incluindo *Y. hibernica*, *Y. rohdei*, *Y. pekkanenii* e o complexo *Y. pseudotuberculosis* (em particular, *Y. pestis*).

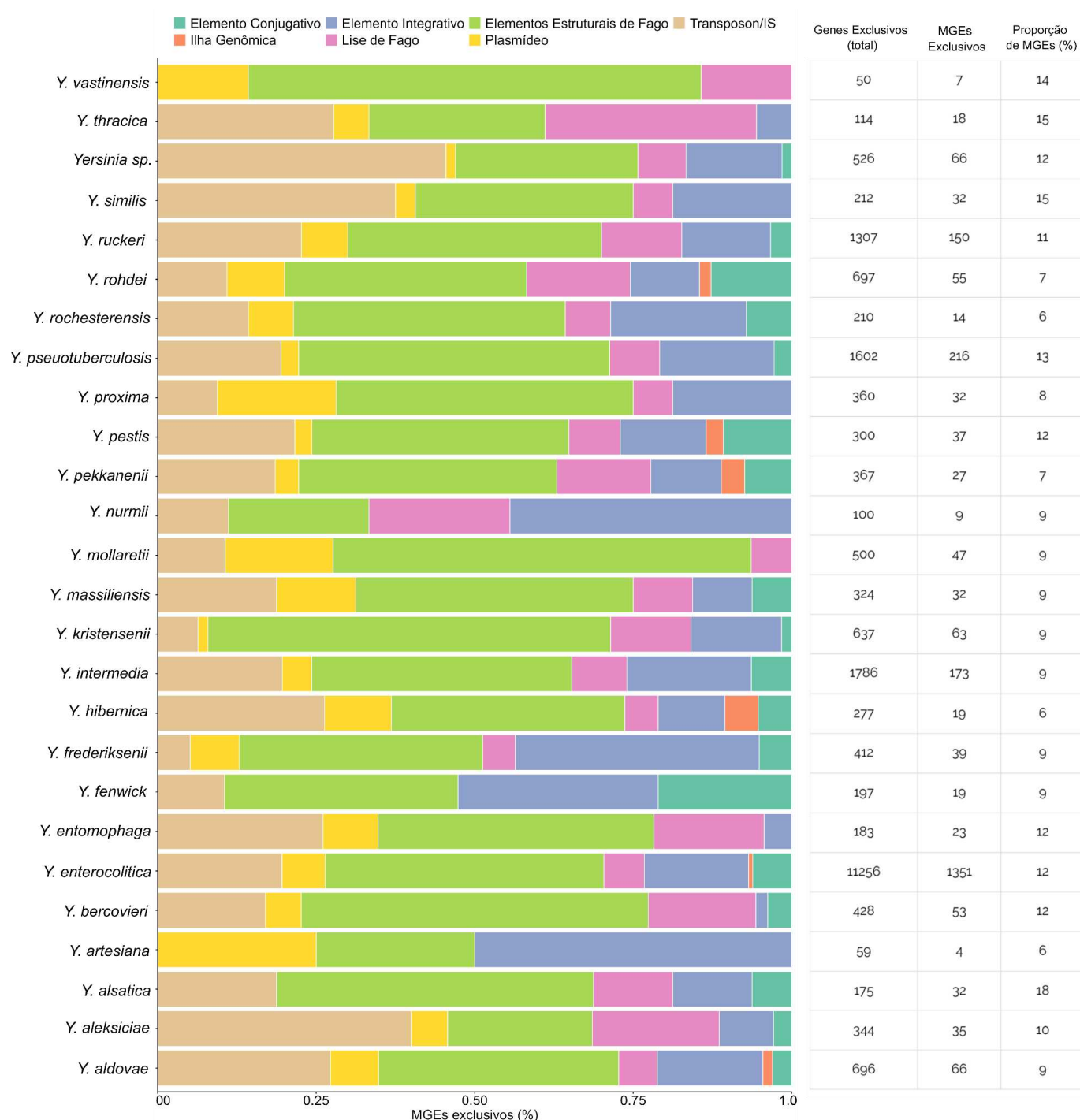


Figura 6. Distribuição dos elementos genéticos móveis (MGEs) entre os genes exclusivos das espécies de *Yersinia*. As barras representam a proporção relativa de genes exclusivos classificados em diferentes categorias de MGEs, incluindo elementos conjugativos, elementos integrativos, componentes estruturais de fago, transposons/sequências de inserção (IS), ilhas genômicas, elementos associados à lise de fago e plasmídeos. À direita, a tabela apresenta o número total de genes exclusivos por cluster e suas respectivas contagens associadas às categorias de MGEs.

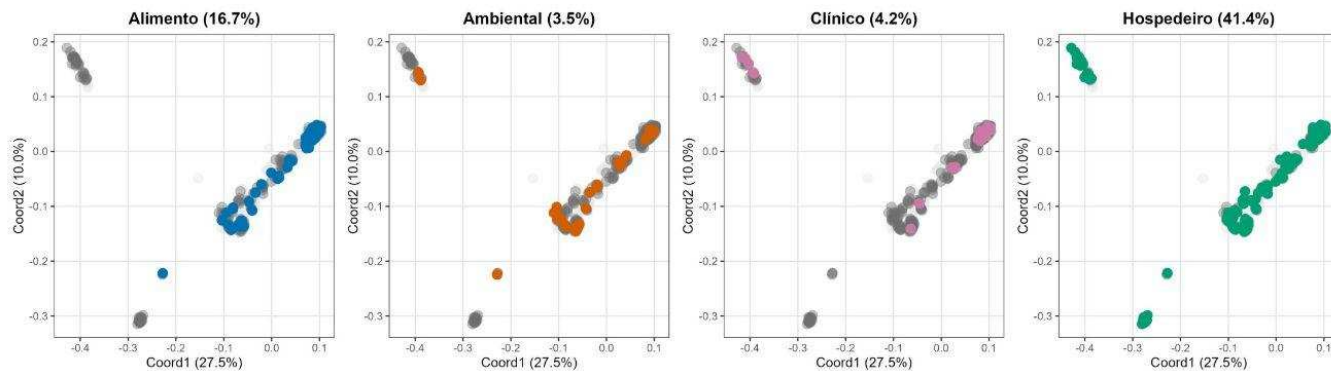
5.5 Associação entre composição gênica e categorias ecológicas dos isolados de *Yersinia*

Dos 1.013 genomas, 670 apresentaram informação classificável para origem do isolamento, 350 para compartimento ecológico e 463 para hospedeiro. A projeção dos metadados ecológicos recuperados do NCBI no espaço de ordenação multivariada baseado na matriz de dissimilaridade do pangenoma revelou padrões sutis de distribuição associados à origem do isolamento, ao compartimento ecológico e ao hospedeiro (Figura 7).

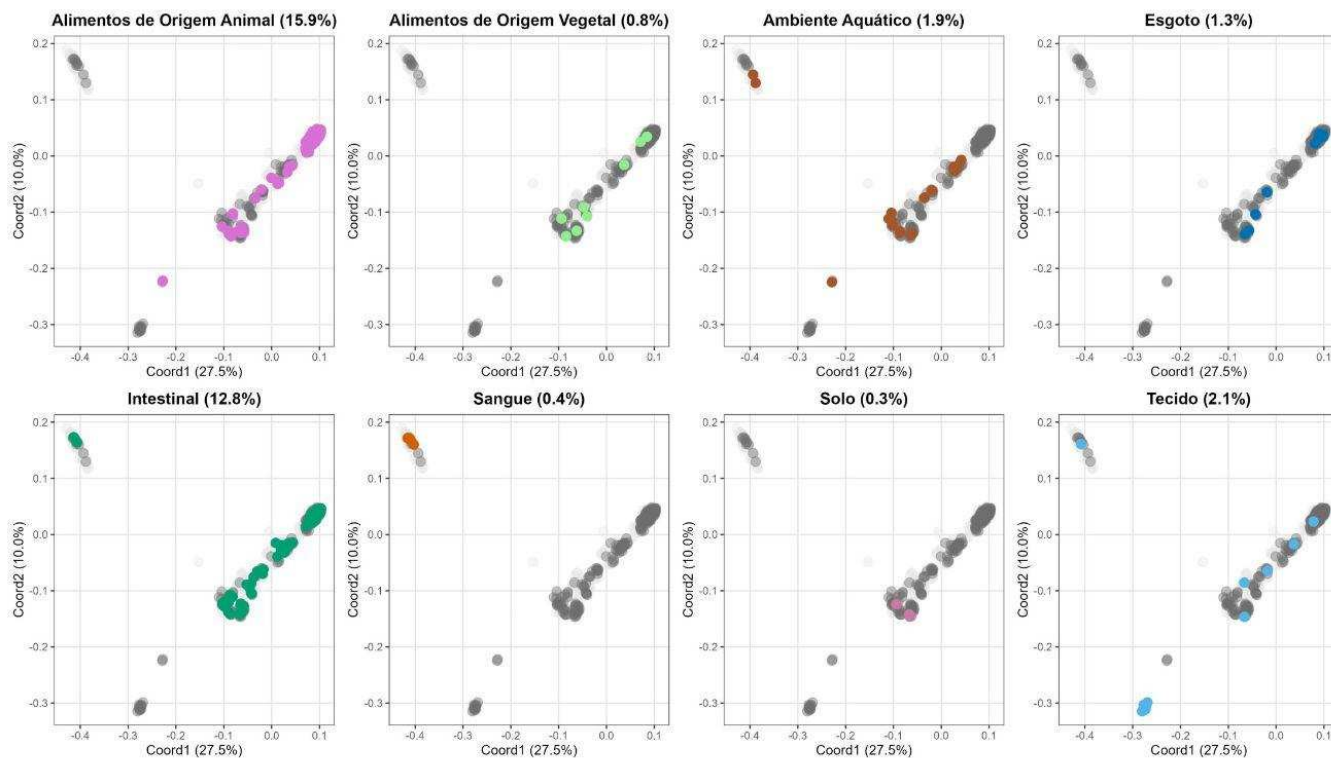
Em relação à classificação dos genomas quanto à origem de isolamento, observou-se a presença de isolados clínicos nos *clusters* (do pangenoma) 1 (complexo *Y. pseudotuberculosis*) e 3 (*Y. enterocolitica*-like). Ainda, considerando a classificação por compartimento ecológico, alguns genomas do *cluster* 1 foram recuperados de amostras de sangue, consistente com a associação dessas linhagens a infecções sistêmicas invasivas na literatura. Por outro lado, isolados provenientes de fontes como ambiente, alimento e hospedeiro apresentaram distribuição ubíqua nos *clusters*, tanto na classificação geral de origem do isolamento (Figura 7a), quanto na classificação detalhada dos compartimentos ecológicos (Figura 7b). Esses padrões refletem a ocorrência de diferentes espécies em fontes de isolamento semelhantes, bem como a recuperação de uma mesma espécie em múltiplos contextos de isolamento.

Por fim, no que diz respeito à estratificação por hospedeiro (Figura 7c), observa-se que o *cluster* formado por *Y. ruckeri*, *Y. entomophaga* e *Y. nurmii* apresentou predominância de isolados associados a peixes e insetos. Em contrapartida, isolados provenientes de humanos e outros mamíferos estão consistentemente distribuídos nos *clusters* 1 (complexo *Y. pseudotuberculosis*) e 3 (que inclui *Y. enterocolitica*) (Figura 7c), estando principalmente associados a amostras recuperadas das fezes e intestino dos hospedeiros (Figura 7b). Notavelmente, observa-se uma ampla distribuição de diversas espécies em hospedeiros mamíferos (incluindo humanos).

a) Origem do Isolado



b) Compartimento ecológico



c) Hospedeiro

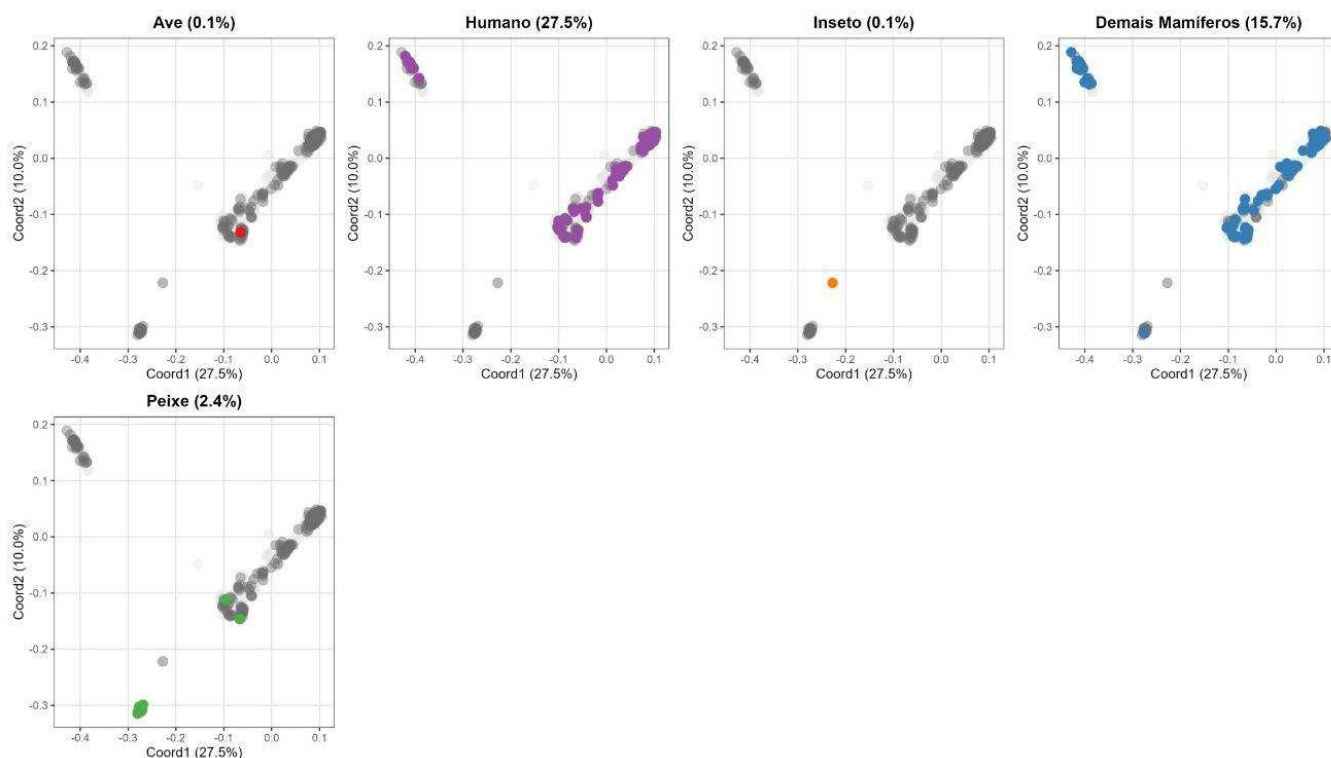


Figura 7. Mapeamento dos metadados ecológicos sobre a estrutura do pangenoma de *Yersinia*. A ordenação foi obtida a partir de um PCoA baseado no conteúdo gênico dos genomas, utilizando a distância de Jaccard calculada a partir da matriz de presença/ausência de genes. Em cada painel, os pontos em cinza claro representam os genomas classificados como “não determinado” para a variável ecológica em questão, enquanto os pontos em cinza escuro correspondem aos genomas pertencentes às demais categorias. Os pontos coloridos destacam, individualmente, a categoria indicada em cada subpainel. As porcentagens entre parênteses indicam a proporção de genomas classificados em cada categoria. (a) Origem do isolado; (b) Compartimento ecológico; (c) Grupo de hospedeiro.

6. DISCUSSÃO

O gênero *Yersinia* inclui espécies amplamente estudadas em função da sua importância clínica, como *Y. pestis* e *Y. enterocolitica*, além de linhagens associadas a animais e a ambientes não clínicos (Bottone, 1999; Wren, 2003; McNally et al., 2016). Nesse contexto, por meio da integração de análises filogenéticas, de similaridade nucleotídica e de pangenoma, o presente estudo caracterizou a diversidade genômica atualmente representada para o gênero e analisou a variação do conteúdo gênico entre linhagens associadas a diferentes contextos ecológicos.

Nesse sentido, a distribuição dos genomas recuperados evidenciou uma forte concentração de registros em *Y. enterocolitica* e *Y. pestis*, enquanto os demais táxons estiveram submostrados (Quadro 1). Esse padrão é consistente com a trajetória histórica dos estudos em *Yersinia*, direcionados a espécies de maior relevância epidemiológica, especialmente os agentes etiológicos da peste e da yersiniose humana (Bottone, 1999; Wren, 2003; McNally et al., 2016). Além disso, a baixa retenção de genomas para *Y. pestis* após a deduplicação é compatível com a reduzida diversidade genômica previamente descrita para a espécie, cuja origem recente a partir de uma linhagem ancestral de *Y. pseudotuberculosis* foi acompanhada por forte estruturação clonal (Achtman et al., 1999; Demeure et al., 2019). Em contrapartida, a maior retenção observada em *Y. enterocolitica* é coerente com a elevada heterogeneidade genética documentada para a espécie, que inclui linhagens associadas a diferentes reservatórios, biotipos e potenciais patogênicos (Reuter et al., 2014; Alenizi et al., 2016). Em resumo, esses resultados demonstram que a desreplicação foi capaz de reduzir o forte desequilíbrio amostral presente nos bancos públicos sem eliminar a diversidade representativa das espécies.

A partir disso, a análise combinada dos agrupamentos genômicos e da reconstrução filogenética revelou uma organização taxonômica consistente, confirmando a delimitação de *Yersinia* em clados nitidamente distintos, a maioria dos quais corresponde a espécies previamente definidas (Figuras 1 e 2) (Savin et al., 2019; Rivas et al., 2026). Todavia, embora a ANI constitua uma métrica importante para delimitar o *status* taxonômico de determinadas linhagens filogenéticas bacterianas, a aplicação de um limiar fixo para definir categorias taxonômicas (por exemplo, espécie) pode não capturar integralmente a estrutura

hierárquica observada em alguns complexos de *Yersinia*, produzindo inconsistências em relação à taxonomia vigente (Rivas *et al.*, 2026). No conjunto analisado, o limiar de ANI $\geq 95\%$ não foi o suficiente para recuperar as subdivisões internas do complexo *Y. pseudotuberculosis* compatíveis com filogenia e com a literatura, refletindo a sensibilidade desse critério em regiões de alta similaridade genômica (Murray *et al.*, 2021). Esse resultado corrobora observações prévias de que a delimitação de linhagens próximas é dependente do *threshold* adotado, incluindo a separação de *Y. similis*, recuperada somente sob limiares mais restritivos ($\geq 96\%$) (Savin *et al.*, 2019). Portanto, a interpretação dos *clusters* baseados em ANI deve ser realizada junto à biologia dos organismos, suas características fenotípicas — incluindo a patogenicidade — bem como sua amplitude filogenética e estrutura interna, de modo a refletir de forma mais consistente a organização taxonômica do táxon em questão.

A reconstrução do pangenoma revelou uma estrutura aberta, fortemente dominada por genes acessórios (95,78% dos genes do pangenoma), indicando elevada diversidade genética (Quadro 2, Figura 3). Embora análises de pangenoma sejam tradicionalmente conduzidas em escala intraespecífica, a abordagem empregada neste estudo teve caráter comparativo e exploratório, buscando descrever a organização do repertório gênico considerando-se a diversidade atualmente representada para *Yersinia* (Medini *et al.*, 2005; Brockhurst *et al.*, 2019). Nesse sentido, a incorporação de organismos mais distantes filogeneticamente não só amplia o repertório gênico total, mas reduz o genoma persistente, o que naturalmente favorece a recuperação de um pangenoma aberto. Logo, a interpretação desse conjunto deve considerar também a diversidade evolutiva e ecológica representada na amostragem. Análises realizadas em outros gêneros bacterianos, como *Flavobacterium* e *Listeria*, nos quais a diversidade interespecífica representa um dos principais fatores associados à expansão do genoma acessório corroboram tal constatação (Kim *et al.*, 2023; Liao *et al.*, 2021). Em *Flavobacterium*, por exemplo, a ocupação de diferentes nichos ecológicos está associada ao acúmulo progressivo de genes acessórios ao longo da diversificação evolutiva do grupo (Kim *et al.*, 2023), enquanto em *Listeria* a integração de múltiplas linhagens e sorotipos resulta em um pangenoma dinâmico, cujo genoma acessório é dominado por elementos genéticos móveis e regiões genômicas variáveis (Liao *et al.*, 2021). Assim, a

elevada proporção de genes acessórios observada em *Yersinia* deve ser interpretada como consequência tanto da diversidade ecológica quanto da amplitude filogenética representada no conjunto analisado.

A análise funcional associada à caracterização do pangenoma tem sido amplamente utilizada para investigar como as categorias funcionais se distribuem entre os genomas *core* e acessórios de um clado (Medini et al., 2005; Tettelin et al., 2008; Tettelin et al., 2012). Nossos resultados demonstram que o genoma persistente de *Yersinia* é composto majoritariamente por genes envolvidos em funções metabólicas, destacando-se processos associados à aquisição e processamento de nutrientes, incluindo aminoácidos, carboidratos e íons inorgânicos (Figura 5). Esse perfil funcional é compatível com a plasticidade metabólica descrita para espécies do gênero *Yersinia*, cuja fisiologia é modulada pela disponibilidade de nutrientes e pelas condições ambientais (Heroven & Dersch, 2014; Chen et al., 2016). Ainda, foram observados genes relacionados à transcrição, tradução, biogênese ribossomal, replicação, recombinação e reparo do DNA, bem como à formação da parede celular e da membrana. Esse conjunto funcional reflete a conservação dos principais componentes da maquinaria celular necessários à manutenção da viabilidade celular, corroborando o papel do *core genome* em funções celulares essenciais (Aggarwal et al., 2022). Assim, neste trabalho, a anotação gênica do genoma persistente fornece uma descrição do repertório funcional compartilhado entre as espécies de *Yersinia* atualmente representadas.

O genoma acessório bacteriano é moldado por processos de aquisição, perda e reorganização gênica ao longo da evolução, nos quais elementos genéticos móveis desempenham papel importante (Brockhurst et al., 2019; Arnold et al., 2022). Sob essa perspectiva, a análise funcional dos genes exclusivos possibilita caracterizar como a fração variável do pangenoma se distribui ao longo da diversidade atualmente representada no gênero *Yersinia*. No presente estudo, constatou-se uma variação expressiva no número de genes exclusivos entre as espécies analisadas, reflexo tanto do número de genomas representados por espécie quanto do grau de diversidade genômica existente entre suas estirpes.

A composição funcional do genoma *cloud* revelou que aproximadamente 6–18% desse conjunto correspondeu a genes associados a elementos genéticos móveis, cuja composição variou entre as espécies. Componentes estruturais de

bacteriófagos constituíram a categoria mais abundante, enquanto transposons, sequências de inserção, elementos integrativos e conjugativos, componentes plasmidiais e ilhas genômicas foram identificados em diferentes proporções (Figura 6). Em bactérias, esses elementos constituem importantes mediadores da transferência horizontal de genes (Arnold et al., 2022). Entretanto, a presença desses elementos não permite inferir diretamente a ocorrência, a intensidade ou o impacto adaptativo de eventos de transferência horizontal nas linhagens analisadas. Em conjunto, esses resultados complementam a caracterização de um pangenoma aberto ao demonstrar que a variabilidade do conteúdo acessório não se restringe ao número de genes exclusivos presentes em cada *cluster*, mas também à composição dessa fração do genoma. Apesar disso, a classificação funcional dos MGEs depende da cobertura e da curadoria das bases de anotação utilizadas, de modo que elementos não caracterizados podem ter permanecido sem identificação. Ainda assim, tal heterogeneidade fornece o contexto para interpretar a estrutura do conteúdo gênico entre os grupos e sua relação com a diversidade ecológica representada no gênero.

Até o momento, os resultados demonstraram que o pangenoma de *Yersinia* apresenta elevada diversidade genética, caracterizada pela expansão do genoma acessório e pela sua heterogeneidade funcional. Entretanto, essas análises não descrevem como essa diversidade se organiza entre os genomas. Nesse contexto, análises de ordenação constituem uma abordagem complementar ao identificar padrões de similaridade baseados na presença e ausência de genes, permitindo avaliar a relação entre o conteúdo gênico, a história evolutiva e a distribuição ecológica das linhagens (Medini et al., 2005; Brockhurst et al., 2019).

A integração entre a análise de ordenação, a filogenia e a similaridade nucleotídica revelou diferentes níveis de organização da diversidade genômica em *Yersinia*. Enquanto a ANI recuperou 25 agrupamentos correspondentes, em sua maioria, às espécies atualmente reconhecidas, a análise de ordenação sintetizou essa diversidade em apenas três *clusters* de conteúdo gênico (Figura 4). Essa diferença demonstra que a elevada plasticidade do genoma acessório, sugerida pelo número de genes exclusivos das espécies, não corresponde à organização global do repertório gênico. Pelo contrário, a distribuição dos genomas no espaço multivariado acompanha a sua estrutura filogenética.

O cluster 1, formado pelo complexo *Y. pseudotuberculosis* (*Y. pestis*, *Y. pseudotuberculosis*, *Y. similis* e *Y. wautersii*), além de constituir um agrupamento monofilético na reconstrução filogenética, apresentou elevada coesividade quanto ao conteúdo gênico (Figura 4). Os metadados recuperados neste estudo demonstraram predominância de isolados clínicos, incluindo genomas provenientes de sangue e fezes, compatíveis com infecções sistêmicas e intestinais (Figuras 2 e 7). Esse padrão é consistente com a literatura, que descreve as espécies desse complexo como os principais representantes patogênicos do gênero, associados predominantemente a mamíferos, incluindo seres humanos (Bottone, 1999; Galindo et al., 2011).

De forma análoga, o cluster 2, formado por *Y. ruckeri*, *Y. entomophaga* e *Y. nurmii*, também foi recuperado como um agrupamento monofilético e apresentou elevado compartilhamento do conteúdo gênico (Figura 4). Os metadados ecológicos indicaram predominância de isolados provenientes de peixes e insetos, em concordância com a ecologia descrita para essas espécies (Figura 7). *Y. ruckeri* é reconhecida como importante patógeno de peixes, enquanto *Y. entomophaga* foi originalmente descrita como patógeno de insetos (Hurst et al., 2011; Wrobel et al., 2019).

Por outro lado, o cluster 3 reuniu as demais espécies do gênero, incluindo aquelas tradicionalmente reconhecidas como pertencentes ao grupo *Y. enterocolitica-like* (Figura 4) (Savin et al., 2019). Embora represente a maior diversidade taxonômica recuperada neste estudo, esse agrupamento manteve elevada similaridade quanto ao conteúdo gênico, apresentando, entretanto, maior dispersão na análise de ordenação em comparação aos demais *clusters*. Esse comportamento acompanha a diversidade filogenética representada no conjunto, indicando que a ampliação da diversidade taxonômica ocorreu sobre um repertório gênico ainda amplamente compartilhado. Os metadados reforçam essa interpretação ao evidenciarem uma distribuição ubíqua de grande parte dessas espécies, com isolados provenientes de ambientes aquáticos, solo, alimentos e diferentes hospedeiros (Figura 7). Essa observação é consistente com descrições clássicas da ecologia de *Y. frederiksenii*, *Y. kristensenii*, *Y. intermedia*, *Y. aldovae*, *Y. mollaretii*, *Y. bercovieri* e *Y. rohdei*, frequentemente recuperadas de múltiplos ambientes, alimentos e hospedeiros (Bottone, 1999; Sulakvelidze, 2000).

Assim, a análise de ordenação demonstra que, embora a expansão do pangenoma tenha sido acompanhada pela incorporação diferencial de genes entre as linhagens, a organização do conteúdo gênico permanece amplamente consistente com a história evolutiva do gênero.

Em resumo, os resultados demonstram que a diversidade genômica de *Yersinia* está estruturalmente organizada de tal modo que a arquitetura taxonômica, a composição do pangenoma e a distribuição do conteúdo gênico refletem, em diferentes escalas, a história evolutiva do gênero. A integração dessas abordagens permitiu caracterizar comparativamente a diversidade atualmente representada para *Yersinia*, buscando fornecer um panorama sistemático da organização genômica do gênero e estabelecer uma base para futuras investigações sobre a evolução e a diversificação de suas linhagens.

7. CONCLUSÃO

Nossos resultados demonstram que o gênero *Yersinia* apresenta elevada diversidade genômica, acompanhada de uma amostragem altamente heterogênea nos bancos de dados disponíveis. Em geral, a organização gênica acompanha a estrutura taxonômica e filogenética do gênero, evidenciando a congruência entre agrupamentos genômicos e os principais clados evolutivos. Além disso, o genoma acessório é predominante e parcialmente representado por elementos genéticos móveis, sugerindo sua contribuição para a variabilidade do repertório gênico. Parte dessa variação parece estar associada aos diferentes contextos ecológicos dos isolados, com padrões de distribuição observáveis nos metadados analisados. Em conjunto, os resultados sustentam a hipótese de que a organização do genoma acessório em *Yersinia* acompanha a estrutura filogenética das linhagens e sugerem que parte dessa organização também está relacionada à heterogeneidade ecológica observada entre os isolados.

8. PERSPECTIVAS FUTURAS

Como perspectivas, os resultados obtidos reforçam a importância do genoma acessório em *Yersinia*, o qual permanece parcialmente explorado. Nesse contexto, a literatura sugere que elementos genéticos móveis desempenham papel central na dinâmica do repertório gênico de algumas espécies do gênero, com evidências de sua participação na aquisição de determinantes de virulência e na modulação do repertório gênico de *Y. pseudotuberculosis* e *Y. enterocolitica*, por exemplo (Moorman and Cohen, 2021). Além disso, a distribuição desses elementos não é homogênea entre as linhagens e pode estar associada a diferenças ecológicas e patogênicas, de modo que sua incorporação em análises comparativas pode ampliar a resolução dos padrões evolutivos em *Yersinia*. Em contrapartida, esse fluxo gênico não é irrestrito, uma vez que sistemas de defesa bacterianos, como o CRISPR-Cas, atuam como barreiras ao fluxo de MGEs, podendo influenciar sua frequência e distribuição entre linhagens, sendo, portanto, relevantes para a interpretação da dinâmica evolutiva do genoma acessório (Faure *et al.*, 2019).

Dado que a diversidade e a distribuição dos MGEs, especialmente em espécies não associadas a doenças em humanos, bem como os mecanismos que regulam sua dinâmica, ainda não estão completamente caracterizados, a compreensão do mobiloma em *Yersinia* permanece incompleta, particularmente no que se refere à extensão da diversidade desses elementos ao longo das diferentes linhagens do gênero. Nesse contexto, a integração entre a dinâmica dos MGEs, a estrutura do pangenoma e os sistemas CRISPR-Cas representa uma abordagem relevante para esclarecer os processos que modulam a diversificação ecológica e evolutiva em *Yersinia*.

REFERÊNCIAS BIBLIOGRÁFICAS

- ACHTMAN, Mark *et al.* *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. ***Proceedings of the National Academy of Sciences***, v. 96, n. 24, p. 14043–14048, 23 nov. 1999.
- AGGARWAL, Sumit Kumar *et al.* Pangenomics in Microbial and Crop Research: Progress, Applications, and Perspectives. ***Genes***, v. 13, n. 4, p. 598, 27 mar. 2022.
- AKENO, Yuya *et al.* A reduced genome decreases the host carrying capacity for foreign DNA. ***Microbial Cell Factories***, v. 13, n. 1, p. 49, 1 dez. 2014.
- ALENIZI, Dhahi *et al.* All *Yersinia enterocolitica* are pathogenic: virulence of phylogroup 1 *Y. enterocolitica* in a *Galleria mellonella* infection model. ***Microbiology***, v. 162, n. 8, p. 1379–1387, 1 ago. 2016.
- AMPHLETT, A. Far East Scarlet-Like Fever: A Review of the Epidemiology, Symptomatology, and Role of Superantigenic Toxin: *Yersinia pseudotuberculosis*-Derived Mitogen A. ***Open Forum Infectious Diseases***, v. 3, n. 1, 1 jan. 2016.
- ARNOLD, Brian J.; HUANG, I-Ting; HANAGE, William P. Horizontal gene transfer and adaptive evolution in bacteria. ***Nature Reviews Microbiology***, v. 20, n. 4, p. 206–218, 12 abr. 2022.
- BOTTONE, Edward J. *Yersinia enterocolitica*: overview and epidemiologic correlates. ***Microbes and Infection***, v. 1, n. 4, p. 323–333, abr. 1999.
- BROCKHURST, Michael A. *et al.* The Ecology and Evolution of Pangenomes. ***Current Biology***, v. 29, n. 20, p. R1094–R1103, out. 2019.
- CARCAUZON, Victoria *et al.* Emergence, global dispersal, and local adaptations of *Yersinia pestis*, the agent of plague. ***Applied and Environmental Microbiology***, v. 92, n. 4, 22 abr. 2026.
- CHEN, Peter E *et al.* Genomic characterization of the *Yersinia* genus. ***Genome Biology***, v. 11, n. 1, p. R1, 4 jan. 2010.
- CHEN, Shiyun; THOMPSON, Karl M.; FRANCIS, Matthew S. Environmental Regulation of *Yersinia* Pathophysiology. ***Frontiers in Cellular and Infection Microbiology***, v. 6, 2 mar. 2016.
- CHKLOVSKI, Alex *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. ***Nature Methods***, v. 20, n. 8, p. 1203–1212, 27 ago. 2023.

CORDERO, Otto X.; POLZ, Martin F. Explaining microbial genomic diversity in light of evolutionary ecology. ***Nature Reviews Microbiology***, v. 12, n. 4, p. 263–273, 3 abr. 2014.

DALZIEL, ANNE C.; ROGERS, SEAN M.; SCHULTE, PATRICIA M. Linking genotypes to phenotypes and fitness: how mechanistic biology can inform molecular ecology. ***Molecular Ecology***, v. 18, n. 24, p. 4997–5017, 16 dez. 2009.

DEMEURE, Christian E. *et al.* *Yersinia pestis* and plague: an updated view on evolution, virulence determinants, immune subversion, vaccination, and diagnostics. ***Genes & Immunity***, v. 20, n. 5, p. 357–370, 3 maio 2019.

DEWAR, Anna E. *et al.* Bacterial lifestyle shapes pangenomes. ***Proceedings of the National Academy of Sciences***, v. 121, n. 21, 21 maio 2024.

DIDELLOT, Xavier; MAIDEN, Martin C.J. Impact of recombination on bacterial evolution. ***Trends in Microbiology***, v. 18, n. 7, p. 315–322, jul. 2010.

DIOP, Awa; RAOULT, Didier; FOURNIER, Pierre-Edouard. Rickettsial genomics and the paradigm of genome reduction associated with increased virulence. ***Microbes and Infection***, v. 20, n. 7–8, p. 401–409, ago. 2018.

FAURE, Guilhem *et al.* CRISPR–Cas in mobile genetic elements: counter-defence and beyond. ***Nature Reviews Microbiology***, v. 17, n. 8, p. 513–525, 5 ago. 2019.

FERRÉS, Ignacio; IRAOLA, Gregorio. An object-oriented framework for evolutionary pangenome analysis. ***Cell Reports Methods***, v. 1, n. 5, p. 100085, set. 2021.

FUKUSHIMA, H.; SHIMIZU, S.; INATSU, Y. *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* Detection in Foods. ***Journal of Pathogens***, v. 2011, p. 1–9, 2011.

GALINDO, Cristi L. *et al.* Pathogenesis of *Y. enterocolitica* and *Y. pseudotuberculosis* in Human Yersiniosis. ***Journal of Pathogens***, v. 2011, p. 1–16, 2011.

GAUTREAU, Guillaume *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. ***PLOS Computational Biology***, v. 16, n. 3, p. e1007732, 19 mar. 2020.

GIL, Rosario *et al.* Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. ***Proceedings of the National Academy of Sciences***, v. 99, n. 7, p. 4454–4458, 2 abr. 2002.

GOLICZ, Agnieszka A. *et al.* Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. ***Trends in Genetics***, v. 36, n. 2, p. 132–145, fev. 2020.

HEROVEN, Ann Kathrin; DERSCH, Petra. Coregulation of host-adapted metabolism and virulence by pathogenic yersiniae. **Frontiers in Cellular and Infection Microbiology**, v. 4, 20 out. 2014.

HURST, Mark R. H. *et al.* *Yersinia entomophaga* sp. nov., isolated from the New Zealand grass grub *Costelytra zealandica*. **International Journal of Systematic and Evolutionary Microbiology**, v. 61, n. 4, p. 844–849, 1 abr. 2011.

JAIN, Chirag *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. **Nature Communications**, v. 9, n. 1, p. 5114, 30 nov. 2018.

KALYAANAMOORTHY, Subha *et al.* ModelFinder: fast model selection for accurate phylogenetic estimates. **Nature Methods**, v. 14, n. 6, p. 587–589, 8 jun. 2017.

KATOH, K.; STANDLEY, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772–780, 1 abr. 2013.

KIM, Minji *et al.* Pangenome analysis provides insights into the genetic diversity, metabolic versatility, and evolution of the genus *Flavobacterium*. **Microbiology Spectrum**, v. 11, n. 5, 17 out. 2023.

LARRALDE, Martin. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. **Journal of Open Source Software**, v. 7, n. 72, p. 4296, 25 abr. 2022.

LE GUERN, Anne-Sophie *et al.* *Yersinia artesianae* sp. nov., *Yersinia proxima* sp. nov., *Yersinia alsatica* sp. nov., *Yersinia vastinensis* sp. nov., *Yersinia thracica* sp. nov. and *Yersinia occitanica* sp. nov., isolated from humans and animals. **International Journal of Systematic and Evolutionary Microbiology**, v. 70, n. 10, p. 5363–5372, 1 out. 2020.

LIAO, Jingqiu *et al.* Nationwide genomic atlas of soil-dwelling *Listeria* reveals effects of selection and population ecology on pangenome evolution. **Nature Microbiology**, v. 6, n. 8, p. 1021–1030, 15 jul. 2021.

MATTHEWS, Chelsea A *et al.* A gentle introduction to pangenomics. **Briefings in Bioinformatics**, v. 25, n. 6, 23 set. 2024.

MCINERNEY, James O.; MCNALLY, Alan; O'CONNELL, Mary J. Why prokaryotes have pangenomes. **Nature Microbiology**, v. 2, n. 4, p. 17040, 28 mar. 2017.

MCNALLY, Alan *et al.* Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. **PLOS Genetics**, v. 12, n. 9, p. e1006280, 12 set. 2016.

MICHOD, Richard E.; BERNSTEIN, Harris; NEDELCU, Aurora M. Adaptive value of sex in microbial pathogens. *Infection, Genetics and Evolution*, v. 8, n. 3, p. 267–285, 10 maio 2008.

MINH, Bui Quang *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, v. 37, n. 5, p. 1530–1534, 1 maio 2020.

MOORMAN, Veronica R.; COHEN, James I. Insights into the individual evolutionary origins of *Yersinia* virulence factor effector proteins. *Plasmid*, v. 114, p. 102562, mar. 2021.

MURRAY, Connor S.; GAO, Yingnan; WU, Martin. Re-evaluating the evidence for a universal genetic boundary among microbial species. *Nature Communications*, v. 12, n. 1, p. 4059, 7 jul. 2021.

NGUYEN, Scott Van *et al.* Comparative genomic insights into *Yersinia hibernica* – a commonly misidentified *Yersinia enterocolitica*-like organism. *Microbial Genomics*, v. 6, n. 9, 1 set. 2020.

PARKHILL, J. *et al.* Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, v. 413, n. 6855, p. 523–527, out. 2001.

PLATT-SAMORAJ, Aleksandra *et al.* The Prevalence of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* in Small Wild Rodents in Poland. *Vector-Borne and Zoonotic Diseases*, v. 20, n. 8, p. 586–592, 1 ago. 2020.

PROKOPENKO, Dmitry *et al.* Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics*, v. 32, n. 9, p. 1366–1372, 1 maio 2016.

REUTER, Sandra *et al.* Parallel independent evolution of pathogenicity within the genus *Yersinia*. *Proceedings of the National Academy of Sciences*, v. 111, n. 18, p. 6768–6773, 6 maio 2014.

RIVAS, Lucia *et al.* *Yersinia fenwicki* sp. nov., isolated from human clinical cases in Aotearoa | New Zealand and Australia. *International Journal of Systematic and Evolutionary Microbiology*, v. 76, n. 1, 22 jan. 2026.

RODRIGUEZ-R, Luis M. *et al.* An ANI gap within bacterial species that advances the definitions of intra-species units. *mBio*, v. 15, n. 1, 16 jan. 2024.

SALAMZADE, Rauf; KOTTAPALLI, Aamuktha; KALAN, Lindsay R. skDER & CiDDER: two scalable approaches for microbial genome dereplication. *bioRxiv: the preprint server for biology*, 6 mar. 2025.

SAYERS, Eric W *et al.* GenBank 2025 update. *Nucleic Acids Research*, v. 53, n. D1, p. D56–D61, 6 jan. 2025.

SAVIN, Cyril *et al.* Genus-wide *Yersinia* core-genome multilocus sequence typing for species identification and strain characterization. ***Microbial Genomics***, v. 5, n. 10, 1 out. 2019.

SULAKVELIDZE, Alexander. *Yersinia* other than *Y. enterocolitica*, *Y. pseudotuberculosis*, and *Y. pestis*: the ignored species. ***Microbes and Infection***, v. 2, n. 5, p. 497–513, abr. 2000.

TETTELIN, Hervé *et al.* Comparative genomics: the bacterial pan-genome. ***Current Opinion in Microbiology***, v. 11, n. 5, p. 472–477, out. 2008.

TETTELIN, Hervé *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”. ***Proceedings of the National Academy of Sciences***, v. 102, n. 39, p. 13950–13955, 27 set. 2005.

TONKIN-HILL, Gerry *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. ***Genome Biology***, v. 21, n. 1, p. 180, 22 dez. 2020.

WANG, He *et al.* Increased plasmid copy number is essential for *Yersinia* T3SS function and virulence. ***Science***, v. 353, n. 6298, p. 492–495, 29 jul. 2016.

WIEDENBECK, Jane; COHAN, Frederick M. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. ***FEMS Microbiology Reviews***, v. 35, n. 5, p. 957–976, set. 2011.

WROBEL, Agnieszka; LEO, Jack C.; LINKE, Dirk. Overcoming Fish Defences: The Virulence Factors of *Yersinia ruckeri*. ***Genes***, v. 10, n. 9, p. 700, 11 set. 2019.

XU, JIANPING. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. ***Molecular Ecology***, v. 15, n. 7, p. 1713–1731, 8 jun. 2006.

YANG, Ruifu *et al.* *Yersinia pestis* and Plague: Some Knowns and Unknowns. ***Zoonoses***, v. 3, n. 1, 2023.

APÊNDICES

APÊNDICE A- Reclassificação taxonômica dos genomas com base no *clustering* por ANI

Acesso do genoma	Classificação do NCBI	Cluster ANI	Atribuição final
GCA_902170665	<i>Yersinia enterocolitica</i>	23	<i>Yersinia proxima</i>
GCA_902170785			
GCA_902170815			
GCA_902170885			
GCA_902170895			
GCA_902170945			
GCA_902170955			
GCA_902170975			
GCA_902170985			
GCA_001135485	<i>Yersinia frederiksenii</i>	4	<i>Yersinia massiliensis</i>
GCA_001151185			
GCA_001218185			
GCA_001222845			
GCA_001319565			
GCA_001319985			
GCA_002192925			
GCA_002591095			
GCA_002192835		18	<i>Yersinia alsatica</i>
GCA_902170305			
GCA_902170295			
GCA_001244565	<i>Yersinia intermedia</i>	4	<i>Yersinia massiliensis</i>
GCA_902170475	<i>Yersinia kristensenii</i>	10	<i>Yersinia rochesterensis</i>
GCA_902170595			
GCA_902170605		13	<i>Yersinia thracica</i>
GCA_902170565			
GCA_030433455	<i>Yersinia sp.</i>	0	<i>Yersinia enterocolitica</i>
GCA_026183365	<i>Yersinia sp.</i>	10	<i>Yersinia rochesterensis</i>
GCA_946902935			
GCA_050586585	<i>Yersinia sp.</i>	17	<i>Yersinia entomophaga</i>

GCA_005937895	<i>Yersinia sp.</i>	15	<i>Yersinia bercovieri</i>
GCA_018224905			
GCA_004100405	<i>Yersinia sp.</i>	19	<i>Yersinia vastinensis</i>

Quadro A1: Reclassificação taxonômica de genomas do gênero *Yersinia* baseada em similaridade nucleotídica.