

# PGAP2: A comprehensive toolkit for prokaryotic pan-genome analysis based on fine-grained feature networks

Received: 28 April 2025

Accepted: 26 September 2025

Published online: 10 November 2025



Congfan Bu<sup>1,2,5</sup>, Hao Zhang<sup>1,2,3,5</sup>, Fengnian Zhang<sup>1,3,4,5</sup>, Wenhao Liang<sup>1,3,4</sup>, Hao Gao<sup>1,2,3</sup>, Jing Zhao<sup>1,2,3</sup>, Fangming Lv<sup>1,2,3</sup>, Ruikun Xue<sup>1,2,3</sup>, Qian Liu<sup>1,4</sup>, Zhewen Zhang<sup>1,2</sup>, Zhong Jin<sup>1,2</sup>✉ & Jingfa Xiao<sup>1,2,3,6</sup>✉

Pan-genome analysis is a crucial method for studying genomic dynamics. By creating pan-genome maps for prokaryotic organisms, we can gain valuable insights into their genetic diversity and ecological adaptability. However, current analytical methods often struggle to balance accuracy and computational efficiency, and they tend to provide primarily qualitative results. This study introduces PGAP2, an integrated software package that simplifies various processes, including data quality control, pan-genome analysis, and result visualization. PGAP2 facilitates the rapid and accurate identification of orthologous and paralogous genes by employing fine-grained feature analysis within constrained regions. Our systematic evaluation with simulated and gold-standard datasets demonstrates that PGAP2 is more precise, robust, and scalable than state-of-the-art tools for large-scale pan-genome data. Furthermore, PGAP2 introduces four quantitative parameters derived from the distances between or within clusters, enabling detailed characterization of homology clusters. Finally, we validate our quantitative findings by applying PGAP2 to construct a pan-genomic profile of 2794 zoonotic *Streptococcus suis* strains. This analysis offers new insights into the genetic diversity of *S. suis*, thereby enhancing our understanding of its genomic structure. PGAP2 is freely available at <https://github.com/bucongfam/PGAP2>.

Prokaryotes, as unicellular microorganisms, demonstrate extraordinary adaptability across diverse ecosystems. This adaptability can largely be attributed to key evolutionary mechanisms such as horizontal gene transfer (HGT), mutations, and genetic drift<sup>1</sup>. These processes are the principal drivers of prokaryotic evolution, continuously introducing novel genetic variations into the gene pool and thus promoting diversity at both the population and species levels<sup>2</sup>. Prokaryotic pan-genome analysis is a systematic method for identifying and characterizing all genes within a specific species. By examining the

distribution patterns and functional diversity of gene families, researchers can gain insights into the evolutionary trajectories and adaptive strategies of species from a population perspective. With ongoing advancements in theoretical frameworks and technological innovations, pan-genome analysis has become a vital approach for studying the genomic dynamics of prokaryotes.

Three key developmental trends are transforming prokaryotic pan-genome research, marking a significant leap forward in our understanding of microbial genomic dynamics. Firstly, the rapid

<sup>1</sup>National Genomics Data Center, China National Center for Bioinformation, Beijing 100101, China. <sup>2</sup>Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China. <sup>3</sup>University of Chinese Academy of Sciences, Beijing 100049, China. <sup>4</sup>Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China. <sup>5</sup>These authors contributed equally: Congfan Bu, Hao Zhang, Fengnian Zhang. <sup>6</sup>These authors jointly supervised this work: Zhong Jin, Jingfa Xiao. ✉ e-mail: [zjin@sccas.cn](mailto:zjin@sccas.cn); [xiaojingfa@big.ac.cn](mailto:xiaojingfa@big.ac.cn)

advancement of sequencing technologies, along with an explosion of large-scale datasets, has revolutionized the field. The number of analyzed strains has grown from just a few dozen in early studies to thousands today<sup>3–5</sup>. This remarkable growth in scale requires innovative analytical methodologies and heightened efficiency in processing the vast amount of data. Secondly, pan-genome analysis is shifting its focus from the localized examination of core genes to a holistic exploration of the entire pan-genome<sup>6–9</sup>. This evolution presents new challenges, such as accurately identifying paralogous genes resulting from recent duplication events and reliably distinguishing shell and cloud gene clusters. Addressing these challenges necessitates advanced analytical techniques capable of navigating this complex landscape. Finally, as we delve deeper into the intricacies of the pan-genome, the scope of research is expanding beyond the simple partitioning of homologous genes. There is now a robust focus on uncovering the evolutionary dynamics of gene families<sup>10–13</sup>. This broader approach highlights the pressing need for comprehensive analytical outputs and enhanced post-inference capabilities to grasp the evolutionary tales these genes truly tell. These collective trends signal a pivotal transformation in pan-genome research, steering us toward more extensive data integration, refined analytical methodologies, and the establishment of robust computational models<sup>14</sup>. Embracing these advancements not only enriches our scientific inquiry but also lays the foundation for groundbreaking discoveries in prokaryotic genomics.

Over the past two decades, a wide variety of pan-genome analysis methods have been developed, broadly categorized into three main categories: reference-based, phylogeny-based, and graph-based. Reference-based methods, such as eggNOG and COG, rely on established orthologous gene databases to identify orthologous genes by aligning genomic sequences with homologous gene annotations in these databases<sup>14–16</sup>. These highly efficient methods are well-suited for analyzing genomes with well-annotated reference data. In contrast, phylogeny-based and graph-based methods employ *de novo* approaches, focusing on gene family homology or positional characteristics. Phylogeny-based methods classify orthologous gene clusters using sequence similarity and phylogenetic information. These methods often employ techniques such as bidirectional best hits (BBH) or phylogeny-based scoring methods. By constructing phylogenetic trees, they aim to reconstruct the evolutionary trajectories of genes<sup>17,18</sup>. Graph-based methods, on the other hand, focus on gene collinearity and the conservation of gene neighborhoods (CGN), which refers to the preservation of gene order across genomes. By creating graph structures to represent relationships across different genomes, these methods enable the rapid and accurate identification of orthologous gene clusters<sup>19–23</sup>.

Current methods for analyzing gene clusters have made progress, but they still face some challenges. Reference-based methods depend on existing annotated datasets, making them less effective for studying new species. They primarily focus on sequence homology and often overlook other structural features, potentially leading to confusion between similar genes or paralogs. Phylogeny-based methods can track the origins of gene duplications through complex phylogenetic trees, but this process can be time-consuming and not ideal for large-scale analyses. Graph-based methods, while computationally efficient, often struggle with accuracy in clustering non-core gene groups, such as mobile genetic elements, and have difficulty adapting to high genomic variability among strains. Most tools primarily provide qualitative descriptions of gene clusters, lacking in quantitative characterization of gene relationships and attributes. This limitation restricts our understanding of orthologous gene functions and their evolution. Overall, there is a clear need for improved methods to enhance our understanding of gene functions and their evolutionary dynamics.

In 2011, we launched the pan-genome analysis pipeline (PGAP), a widely used workflow for analyzing genomes based on gene

homology, designed for dozens of strains<sup>16</sup>. With the dramatic increase in the number of prokaryotic genomes, there is a higher demand for pan-genome analysis.

Now, we present PGAP2, an enhanced method for pan-genome analysis that accommodates thousands of genomes. It offers comprehensive workflows and visualization tools to effectively help users interpret input strain properties. Validation with simulated and carefully curated datasets shows that PGAP2 consistently outperforms other methods' stability and robustness, even under genomic diversity. We analyzed the pan-genome of 2794 zoonotic *Streptococcus suis* strains to showcase its real-world application. This study highlights PGAP2's effectiveness in handling diverse prokaryotic populations and its potential to advance prokaryotic genomics research.

## Results

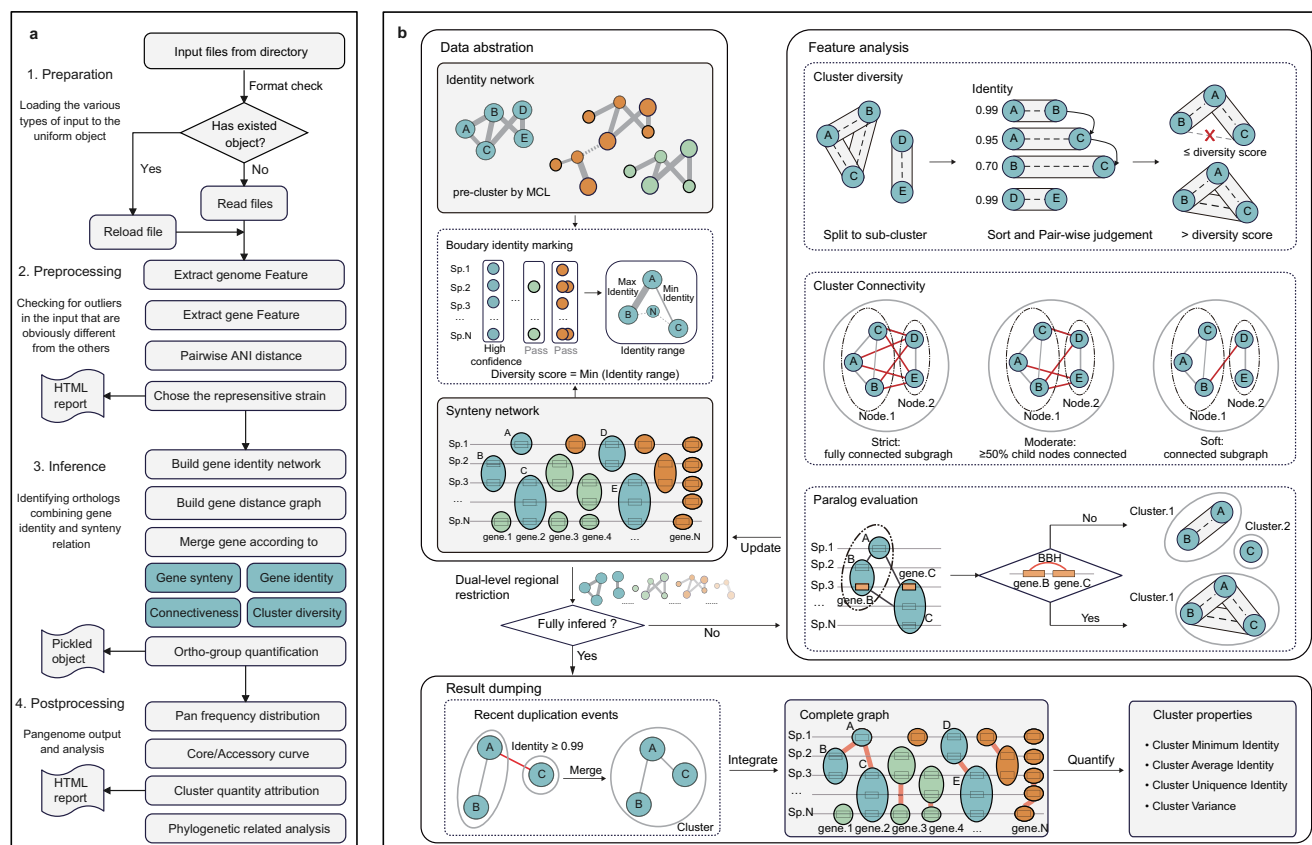
### Overview of the method

The workflow of PGAP2 can be broadly divided into four successive steps: data reading, quality control, homologous gene partitioning, and postprocessing analysis (Fig. 1a):

**PGAP2 is compatible with various input formats.** PGAP2 accepts four types of input data: GFF3, genome FASTA, GBFF and GFF3 with annotations and genomic sequences. The last one, produced by genome annotation tools such as Prokka, combines a GFF3 annotation with its corresponding nucleotide sequence. PGAP2 can accept a mix of different input formats and identify the input format based on the file suffix. After reading and validating all the data, PGAP2 organizes the input into a structured binary file to facilitate checkpointed execution and downstream analysis.

**Perform quality control and generate feature visualization reports.** If no specific strain is designated, PGAP2 will select a representative genome based on gene similarity across strains. Briefly, PGAP2 evaluates outliers using two methods. The first method is based on the similarity of average nucleotide identity (ANI). If a strain's similarity to the representative genome falls below a certain threshold (e.g., 95%), it is classified as an outlier. The second method compares the number of unique genes in the strain with those in other strains. A strain with a higher number of unique genes is more likely to be classified as an outlier. Additionally, PGAP2 generates interactive HTML and vector plots to visualize features such as codon usage, genome composition, gene count, and gene completeness, helping users assess input data quality (Supplementary Figs. 1 and 2).

**Infer orthologs through fine-grained feature analysis under a dual-level regional restriction strategy.** The process of inferring orthologous genes in PGAP2 can be divided into three key steps: data abstraction, feature analysis, and result dumping (Fig. 1b). Essentially, PGAP2 organizes the required data into two distinct networks: a gene identity network and a gene synteny network. In the identity network, edges represent the degree of similarity between genes, while in the synteny network, edges denote adjacent genes - specifically, genes that are one position apart. PGAP2 then splits gene clusters that contain redundant genes within the same strain, using conserved gene neighbor (CGN) to ensure that the graph remains acyclic (Supplementary Note 1). Finally, PGAP2 calculates a diversity score using the updated networks. This score helps evaluate the conservation level of orthologous genes in subsequent analyses. In the second step, PGAP2 performs orthologous gene inference by traversing all subgraphs in the identity network. This process is divided into two main stages: regional refinement and feature analysis. In each iteration, PGAP2 applies a dual-level regional restriction strategy, evaluating gene clusters only within a predefined identity and synteny range. This targeted approach significantly reduces search complexity by focusing on a confined radius, enabling more detailed and efficient analysis of



**Fig. 1 | Overview of the PGAP2 workflow. a** The flowchart depicts four sequential stages of the PGAP2 pipeline, arranged from top to bottom: data preprocessing, quality control, gene clustering, and postprocessing analysis. **b** The core algorithm of PGAP2 begins by constructing an identity network and a synteny map, which

serve as the foundational data structures for ortholog inference. Following regional refinement, PGAP2 iteratively merges nodes based on gene cluster diversity, connectivity, and Bidirectional Best Hit (BBH) criteria. The abbreviation “sp.” in this figure refers to “species”.

features within these clusters (Supplementary Note 2). The reliability of orthologous gene clusters is evaluated using three criteria: 1) gene diversity, 2) gene connectivity, and 3) the bidirectional best hit (BBH) criterion, which is applied to duplicate genes within the same strain. The detailed algorithms used for these assessments are outlined in the methods section and Supplementary Note 3. Once merged, gene clusters are updated in the synteny network, which undergoes further iterations until all clusters no longer meet the established criteria. Finally, PGAP2 merges nodes with exceptionally high sequence identity, which often arise from recent duplication events driven by horizontal gene transfer or insertion sequences. The properties of the orthologous gene clusters are then outputted, including the average identity, minimum identity, average variance, and uniqueness to other clusters, providing insights into genome dynamics. More detail of the graph algorithm is described in Supplementary Methods.

#### Record pan-genome profile and generate visualization reports.

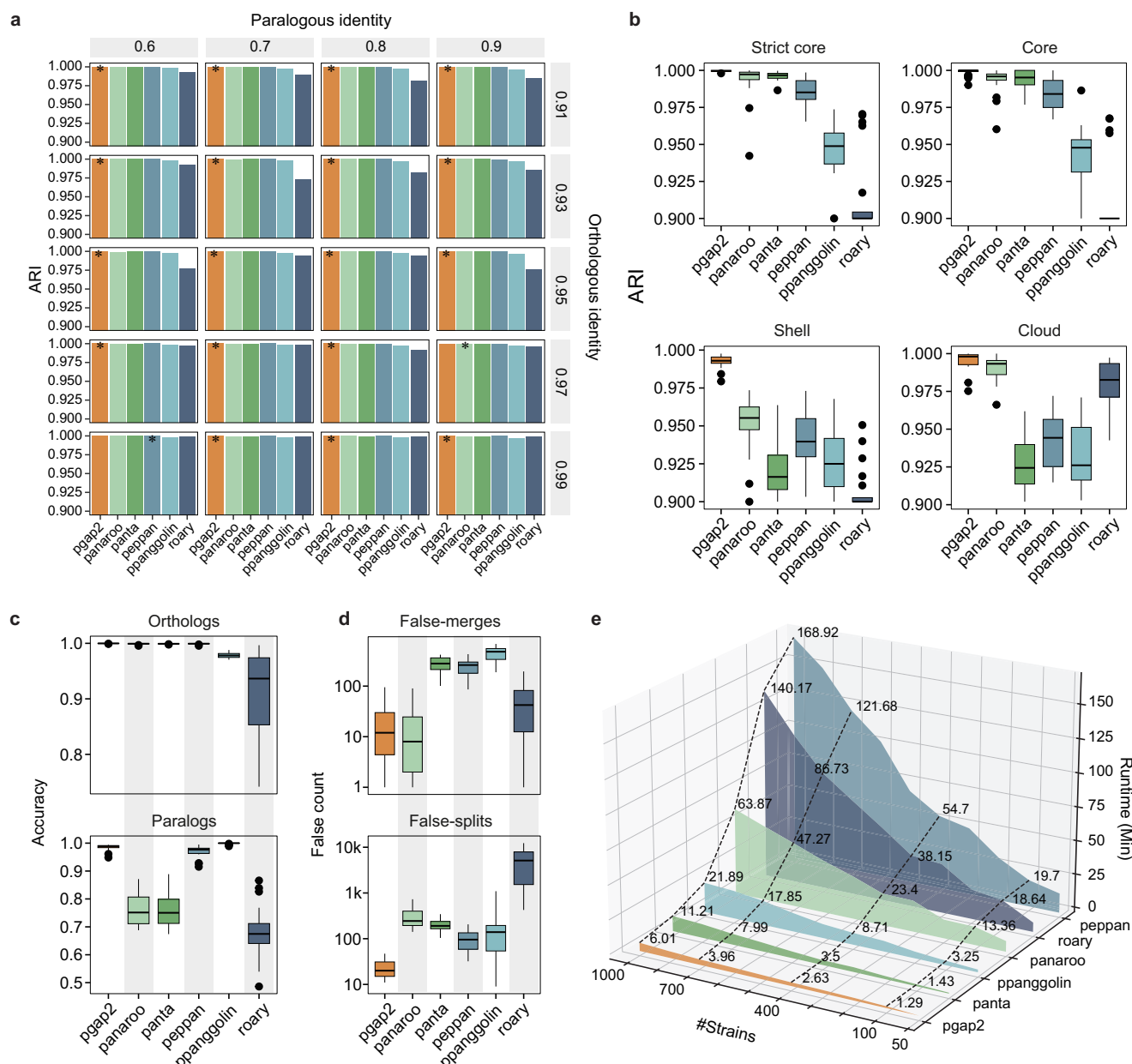
Similar to preprocessing, the postprocessing module generates interactive visualizations in both HTML and vector formats. These visualizations display the rarefaction curve, statistics of homologous gene clusters, and quantitative results of orthologous gene clusters. PGAP2 employs the distance-guided (DG) construction algorithm, initially proposed in PanGP<sup>13</sup>, to construct the pan-genome profile. Finally, by integrating multiple software tools, PGAP2 provides comprehensive workflows, which include sequence extraction, single-copy phylogenetic tree construction, and bacterial population clustering. This integration ensures a seamless experience for users (Supplementary Figs. 3–6).

#### Methods' performance on the simulated dataset

To assess PGAP2's robustness, we evaluated its accuracy using different thresholds for orthologs and paralogs, simulating variations in species diversity. We compared PGAP2 against five state-of-the-art tools, including Roary, Panaroo, PanTa, PPanGGOLiN, and PEPPAN, all tested using their respective default parameters. The thresholds for orthologs were adjusted from 0.99 to 0.91, while those for paralogs ranged from 0.9 to 0.6. This approach resulted in 16 combinations of datasets.

We used the adjusted rand index (ARI) to evaluate the accuracy of software tools by indicating the differences between two matrices. The results show that all the software tools achieved an average accuracy of over 90%. PGAP2 consistently outperformed the other tools in most gradient combinations (14 out of 16), achieving an impressive average accuracy of 0.9997. This exceeds that of the second-best tool, PEPPAN, which recorded an accuracy of 0.9990 (Fig. 2a and Supplementary Table 1). These results illustrate PGAP2's robustness in simulated extreme diversity scenarios, demonstrating its ability to maintain high accuracy even under challenging conditions. These findings suggest that PGAP2 is particularly well-suited for analyzing highly diverse strains, where high precision is critical for accurate downstream analyses and biological interpretation. Furthermore, robustness tests using simulated incomplete annotations demonstrated that PGAP2 maintained over 99% clustering accuracy, even with significant gene loss and fragmentation. This further validates its reliability across various low-quality input conditions (Supplementary Fig. 7).

By categorizing gene clusters from different perspectives, we further evaluated the accuracy (or error rate) of each software from



**Fig. 2 | Performance comparison of various software on simulated data.** **a** The accuracy of PGAP2 and other software was assessed using the Adjusted Rand Index (ARI). The dataset comprises 16 combinations formed by varying orthologous thresholds (from 0.99 to 0.91) and paralogous thresholds (from 0.9 to 0.6). The software with the highest accuracy is marked with an asterisk (\*). **b** shows the accuracy of each software across different cluster categories. **c** presents the accuracy of PGAP2 on orthologous and paralogous gene clusters, respectively. **d** summarizes the error types generated by each software: False merges refer to the

incorrect merging of multiple clusters, and False splits refer to the incorrect splitting of a single cluster. **e** provides a runtime comparison of each software across varying numbers of genomes (from 50 to 1000), tested with 8 threads and default parameters. For panels **b–d**, The boxplots represent the first, second (median), and third quartiles; whiskers represent the most extreme data points within 1.5× IQR of the quartiles; and black dots indicate outliers. All sample sizes are  $n = 16$ , corresponding to the datasets described in panel (a).

three dimensions to demonstrate that PGAP2 does not exhibit any specific error bias.

First, we assessed the accuracy of PGAP2 across different categories of gene families, including strict core, core, shell, and cloud (Fig. 2b, Supplementary Table 2). Here, strict core genes are defined as those present in 100% of genomes; core genes are present in  $\geq 95\%$  of genomes; shell genes are found in 15% to  $< 95\%$  of genomes; and cloud genes are rare, occurring in  $< 15\%$  of genomes. The results showed that PGAP2 achieved the highest accuracy across all gene family types, with an average accuracy of  $99.62\% \pm 0.32\%$ . Notably, other software tools showed significantly lower accuracy in shell gene families, likely due to

the high diversity in shell clusters. In contrast, PGAP2 maintained an error rate of only 0.77% in shell gene families, significantly outperforming the others, which ranged from 4.96% to 30.39%. This improvement can be attributed to PGAP2's dynamic parameter planning, which enables it to maintain high accuracy even in highly diverse gene families.

Second, we classified gene families based on their evolutionary origin into orthologous and paralogous types. As expected, nearly all software tools achieved high accuracy in identifying orthologous genes, ranging from 90.82% to 99.91% on average, with PGAP2 reaching the highest at 99.97%. However, notable differences were observed



in the identification of paralogous genes. PGAP2, PEPPAN, and PPanGGOLiN performed significantly better than the remaining three tools, which can be explained by their respective optimizations: PEP-PAN uses phylogenetic tree-based methods, PPanGGOLiN applies maximum likelihood estimation, and PGAP2 incorporates dynamic parameter planning (Fig. 2c, Supplementary Tables 3–4).

Finally, we analyzed the types of errors made by each software and classified them into two categories: false merges, where two or more clusters are incorrectly merged into a single cluster, and false splits, where a single cluster is incorrectly split into multiple clusters. The results revealed that the main error types varied across software. For example, the primary errors in Panaroo and Roary were false splits, while PanTa, PEPPAN, and PPanGGOLiN were more prone to false merges. PGAP2 showed mean counts of 20.1 for false merges and 23 for false splits, indicating no significant error bias. Interestingly, although PPanGGOLiN achieved the highest accuracy in identifying paralogous clusters, it appeared to cause more false merges in orthologous gene clusters, suggesting a bias toward this type of error (Fig. 2d, Supplementary Tables 5–6).

### Computation time

We generated 11 additional gradient-simulated datasets to assess the runtime of PGAP2 and competing tools. The number of genomes ranged from 50 to 1000, with an average of 4500 coding genes per genome. All software was tested using 8 threads with default parameters (Fig. 2e, Supplementary Tables 7–8). PGAP2 demonstrated near-linear scalability, with no significant computational bottlenecks as the dataset size increased. On the largest dataset, PGAP2 completed the analysis in approximately 6 min, nearly twice as fast as the second-fastest tool, PanTa, and 23 times faster than Roary, while requiring only one-third of the memory used by Roary. These findings suggest that PGAP2 maintains high accuracy and is well-equipped to handle the demands of large-scale prokaryotic pan-genome analyses.

To further evaluate the scalability of PGAP2 with real-world datasets, we chose three prokaryotic species with the most complete genomes available in the NCBI RefSeq database: *Klebsiella pneumoniae* (25,437 genomes), *Staphylococcus aureus* (18,411 genomes), and *Escherichia coli* (44,284 genomes). Using 32 threads and default settings, PGAP2 successfully conducted pan-genome analyses for all three datasets. The runtimes were about 6 h for *S. aureus*, 12 h for *K. pneumoniae*, and 119 h for *E. coli*, with peak memory usage reaching up to 886 GB. These results demonstrate that PGAP2 can handle some of the largest collections of prokaryotic genomes currently available (Supplementary Fig. 8).

### Benchmark datasets construction for pan-genome analysis

To further assess the accuracy of PGAP2 using real-world data, we manually created a benchmark pan-genome dataset composed of 50 strains of *Escherichia coli*, representing eight major phylogroups (Supplementary Table 9). The workflow for constructing this dataset is illustrated in Fig. 3a, and detailed quality control criteria are provided in the methods section. Briefly, we selected representative strains based on their genome distances. Raw annotations were manually curated and merged based on semantic similarity, while homologous gene families were defined by integrating functional annotations from SwissProt. Homologs were further refined using synteny information, followed by phylogenetic tree construction and BBH-based paralog identification. When comparing the raw cluster dataset to the benchmark dataset, several improvements have been made to boost accuracy. The benchmark dataset has removed pseudogenes and short sequences, corrected clusters that needed merging or splitting, and resolved all instances of paralogous genes. The Cohen's kappa score between the raw and benchmark datasets is 0.6713 (Supplementary Table 10), indicating a substantial level of agreement and reflecting the changes made through manual curation. Ultimately, the benchmark

dataset includes 224,798 genes organized into 14,360 homologous gene clusters, which consist of 3063 core gene families, 2494 shell gene families, and 8803 cloud gene families (Supplementary Table 11 and Supplementary Data 1).

The pan-genome profile illustrates the open nature of the *E. coli* pan-genome (Supplementary Fig. 3C). Core genes are mainly associated with essential conserved functions, including translation, ribosomal structure, biogenesis, and energy production and conversion. Shell genes are involved in functions related to transcription, cell motility, intracellular transport, and secretion, while cloud genes are enriched in functions related to or associated with transcription, replication, recombination, and repair, with a significant proportion of genes of unknown function (Supplementary Fig. 10). These functional patterns are consistent with previous observations<sup>24–26</sup>.

We constructed a phylogenetic tree using curated single-copy core clusters, which revealed that all strains grouped into their respective phylogroups (Supplementary Fig. 11). To further validate the reliability of our dataset, we examined it from four perspectives: the sequence similarity among proteins within the clusters (Fig. 3b), the consistency of functional annotations (Fig. 3c), the average genetic distance among genes within the clusters (Fig. 3d), and the difference in genetic distances between intra-cluster genes and the nearest inter-cluster genes (Fig. 3e). Our results demonstrated that the average protein sequence similarity within all clusters exceeded 70%, and the consistency of functional annotations for the annotated gene families was also above 70%. Furthermore, the average genetic distance among genes within clusters was less than 0.3, which is significantly smaller than the distance to the nearest neighboring cluster (two-sided Wilcoxon signed-rank test,  $p < 2.22 \times 10^{-16}$ ). These findings highlight the accuracy and reliability of our curated dataset.

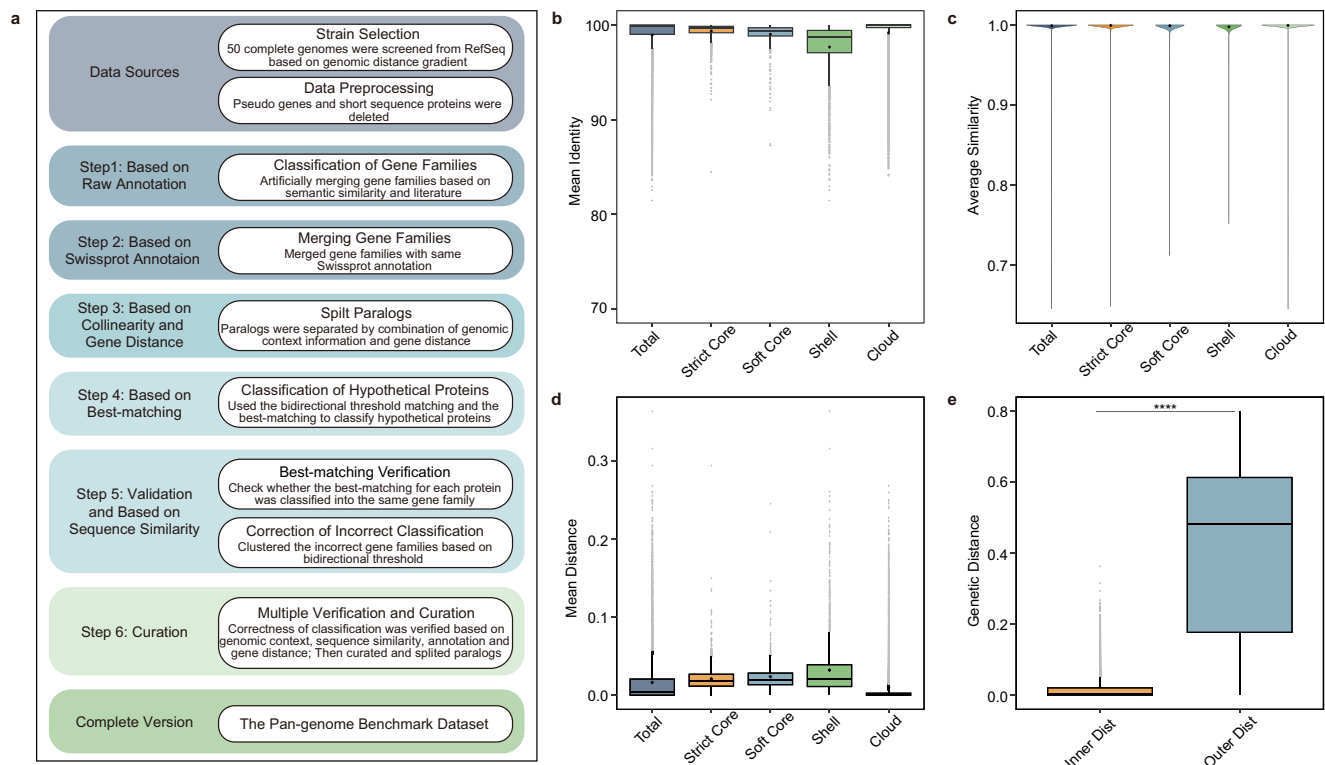
### Methods' performance on the *Escherichia coli* data dataset

The performance of each pan-genome analysis method on the curated *Escherichia coli* dataset is illustrated in Fig. 4a and c. All methods achieved an accuracy of over 0.98. We evaluated three sensitivity modes in PGAP2 - soft, mid, and strict, which define the connectivity thresholds used in gene clustering. The strict mode requires a fully connected graph, wherein all gene pairs must exceed the homology threshold. This mode is suitable for identifying conserved genes across different strains. In contrast, the soft mode only requires minimal connectivity (one edge), making it more appropriate for identifying highly variable homologous gene clusters across species. The testing of these modes yielded accuracies of 0.9932, 0.9933, and 0.9932, respectively, all of which exceed the accuracy of the second-best method, Panaroo, which achieved an accuracy of 0.9917 (Fig. 4a and Supplementary Table 12). Regarding core orthologous gene clusters, Fig. 4c shows that Panaroo identified the highest number, with 3072 clusters, followed closely by PPanGGOLiN with 3064 and PGAP2 with 3063. Among these methods, PGAP2's result most closely matched the gold standard. Conversely, PanTa identified the fewest clusters, totaling 2900 (Supplementary Table 13).

We also examined misclassified genes and gene clusters (Fig. 4a). Depending on the sensitivity mode used, PGAP2 misclassified approximately 1500 clusters, significantly fewer than the average of  $3,321.0 \pm 1,512.18$  clusters misclassified by other tools. PGAP2 misclassified only six strict core clusters, compared to ten identified by Panaroo. This is notably lower than the overall average of  $79.8 \pm 38.92$  clusters (Supplementary Table 14).

### Methods' performance on the *Salmonella enterica* benchmark dataset

To address potential biases related to the dataset or strain preferences, we re-evaluated PGAP2 using a different gold standard dataset that included 15 *Salmonella enterica* strains (Fig. 4b, d, and Supplementary Table 15). Similar to its performance on the *E. coli* benchmark, all



**Fig. 3 | The data profiling of the *Escherichia coli* benchmark.** **a** The construction of the benchmark includes six steps: merging clusters based on annotations (Steps 1 and 2), splitting paralogs based on collinearity (Step 3), classifying and correcting clusters based on best-matching (Steps 4 and 5), and multiple verifications and manual curation to generate the complete version (Step 6). **b** The average protein sequence similarity across various types of clusters, with sample sizes:  $n_{\text{total}} = 14,360$ ,  $n_{\text{strict core}} = 2331$ ,  $n_{\text{soft core}} = 732$ ,  $n_{\text{shell}} = 2494$ ,  $n_{\text{cloud}} = 8803$ . **c** The average semantic similarity of domain annotations across various types of clusters. The violin plots show the kernel density estimate of the semantic similarity distribution, where the width represents the relative frequency at the type of clusters, diamond symbols represent mean. The sample size only includes the clusters with annotations:  $n_{\text{total}} = 9872$ ,  $n_{\text{strict core}} = 2250$ ,  $n_{\text{soft core}} = 672$ ,  $n_{\text{shell}} = 2011$ .

methods demonstrated an overall accuracy exceeding 0.98. PGAP2 achieved the highest accuracy at 0.9964, surpassing the second-best method, Roary, which reached an accuracy of 0.9949 (Supplementary Tables 16–17).

Notably, Roary and PanTa performed comparatively better on *Salmonella* but showed lower performance on *Escherichia coli*, suggesting that these methods may have species-specific thresholds or parameter sensitivities. This observation is consistent with their behavior on simulated datasets and highlights a potential bias in these tools when applied across diverse taxa. Regarding error rates, PGAP2 yielded the fewest misclassified genes and gene clusters. Specifically, it reported 1854 misclassified genes, compared to 2203 for Roary. PGAP2 also misclassified only 425 gene clusters, while PEPPAN and PPanGOLiN misclassified 489 clusters each (Supplementary Table 18).

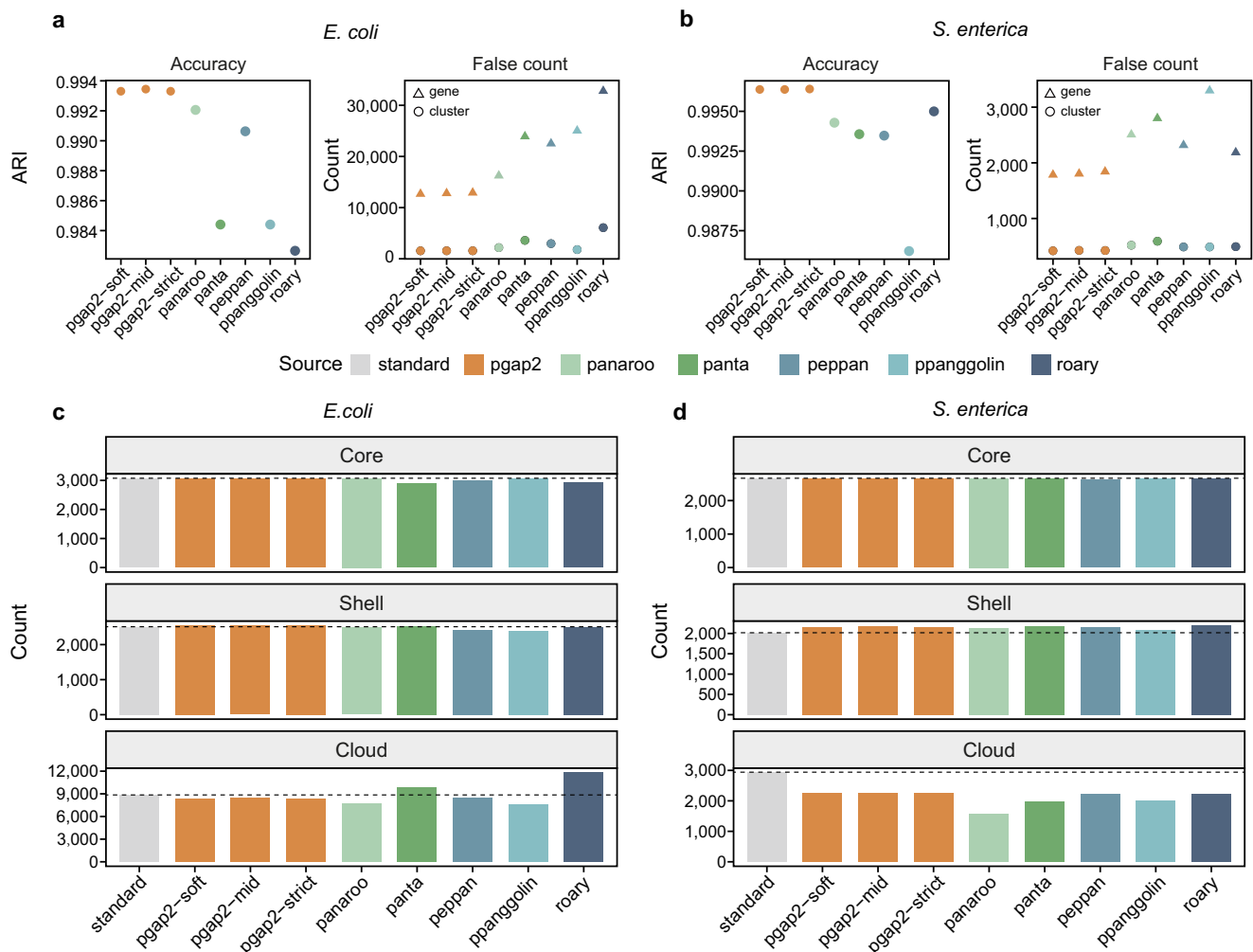
Taken together, our results demonstrate that PGAP2 maintains high accuracy across species and does not exhibit species-specific performance bias. Its consistent superiority across independent benchmarks confirms its robustness and reliability for pan-genome inference in diverse prokaryotic datasets.

### A pan-genome analysis for *Streptococcus suis*

PGAP2 is capable of rapidly generating high-quality pan-genome profiles from thousands of strains. To demonstrate its scalability and robustness on a high-volume dataset, we analyzed all 2794 strains of

*Streptococcus suis*, a zoonotic pathogen known for its extensive genetic diversity and large gene pool, using data available from the NCBI RefSeq database.

PGAP2 completed the pan-genome analysis of 2794 *S. suis* strains within 20 minutes using 32 threads. A total of 17,646 orthologous gene clusters were identified, comprising 1395 core genes (7.90%), 6.98% shell genes, and 85.12% cloud genes (Fig. 5a). This number is smaller than the 29,738 gene clusters identified in 1204 *S. suis* strains using Roary<sup>27</sup>. This discrepancy may stem from Roary's clustering strategy, which is prone to over-segment gene clusters with high sequence diversity, especially those annotated as shell or cloud genes, thereby introducing false-split errors, as demonstrated in our benchmarking evaluation (Fig. 2d). Core genes represented an average of 70.67% of all coding genes in each strain, slightly higher than previous studies, such as 66.5% by Zhang et al.<sup>28</sup> and 63% by Xia et al.<sup>29</sup>. Among the analyzed strains, *S. suis* YB51 Colony29 exhibited the highest proportion of core genes at 85.42%, while *S. suis* SS/UPM/MY/F001 had the lowest at 55.95%. These differences are likely due to the varying number of coding genes present in their genomes (Supplementary Fig. 12). The pan-genome growth curve follows a sub-linear power-law trend, indicating that as more genomes are included, new genes continue to be discovered but at a diminishing rate. The positive exponent confirms an expanding pan-genome (Fig. 5b), supporting the idea of an open pan-genome, which is consistent with previous studies<sup>27–30</sup>.



**Fig. 4 | Comparison of pan-genome predictions with curated datasets. a** The accuracy and false counts for each software were evaluated using a manually curated dataset consisting of 50 *E. coli* genomes. **b** The accuracy and false counts for each software evaluated using a manually curated dataset of 15 *Salmonella*

genomes. **c** The distribution of clusters within pan-genome groups derived from the *E. coli* dataset. **d** The distribution of clusters within pan-genome groups derived from the *Salmonella* dataset. The terms “soft”, “mid”, and “strict” refer to the different levels of clustering strictness employed in PGAP2.

### Quantitative attributes of *Streptococcus suis* pan-genome profile

PGAP2 provides quantitative metrics for each gene cluster based on four parameters: Uniqueness (Uni), Minimum identity (Min), Mean identity (Mean), and Identity variance (Var) (Fig. 5c). Uni describes inter-cluster relationships, whereas Min, Mean, and Var reflect intra-cluster sequence similarity. Uni quantifies the minimum distance between an ortholog and other clusters, with lower values indicating greater separation. When Uni exceeds a given threshold (0.7 by default), it defines a unique cluster, indicating that there is insufficient homology for classification within an established gene family. Our analysis revealed that over two-thirds of homologous gene clusters are unique, with most originating from core and cloud clusters. In contrast, shell clusters tend to exhibit higher homology (Fig. 5d). This finding suggests that cloud genes are predominantly influenced by horizontal gene transfer in *S. suis*, which contributes to its genetic diversity.

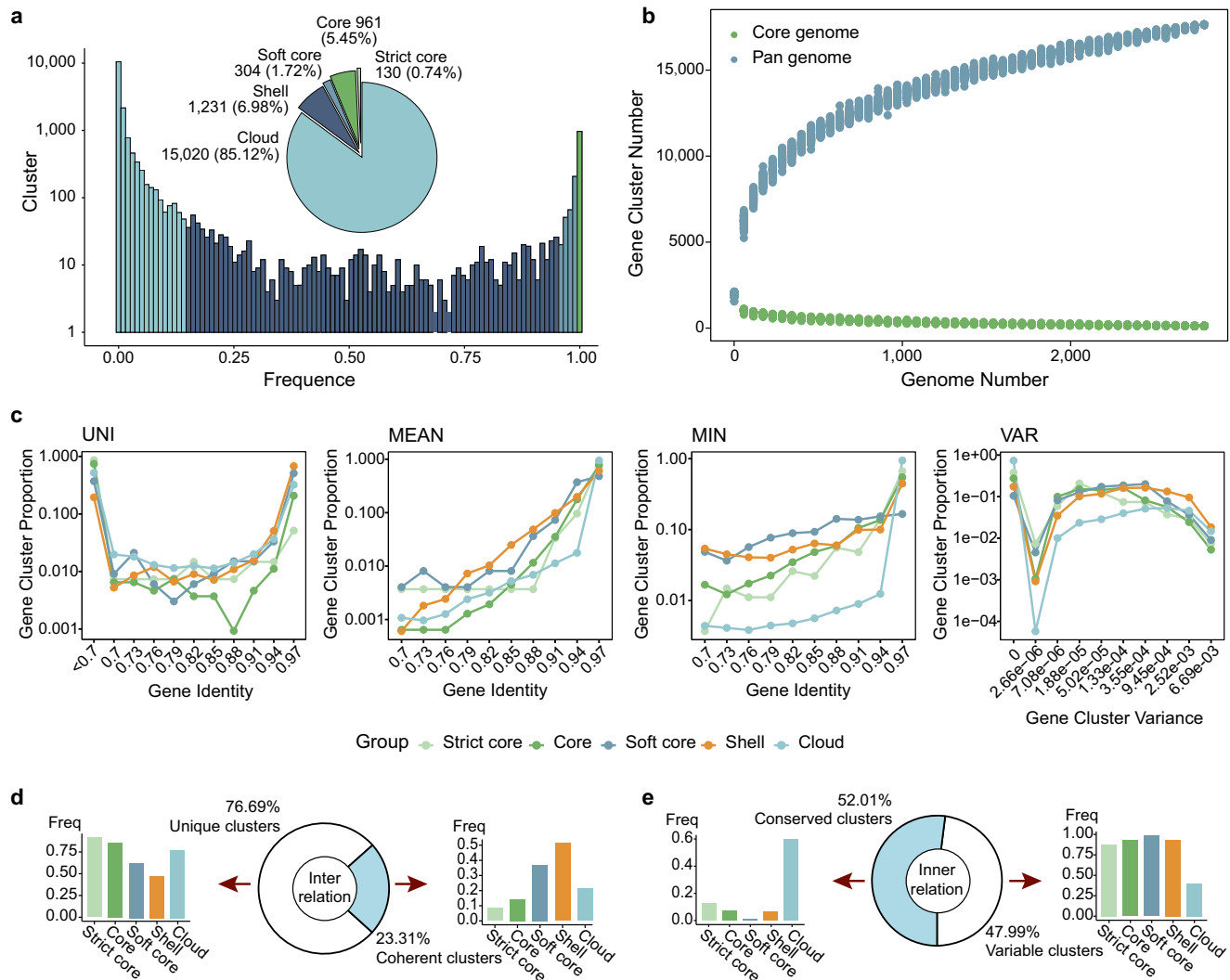
Min, Mean, and Var are key to classifying gene clusters into conserved and variable categories. Conserved clusters, where both Min and Mean exceed a threshold, show high sequence consistency, due to purifying selection (Fig. 5e). In contrast, variable clusters exhibit increased sequence diversity, particularly when either Min or Mean falls below the threshold. This variation may arise from evolutionary

forces such as genetic drift and mutations. Notably, soft core and shell genes demonstrate the greatest homolog diversity, likely owing to their stability within populations, making them more susceptible to environmental selection pressures (Supplementary Fig. 13).

The enrichment analysis of high-variance clusters indicated that these homologs are preferentially associated with pathways involved in environmental interactions (Supplementary Table 19). The most significant enrichment was observed in quorum sensing, a bacterial communication mechanism that allows cells to sense population density via autoinducer molecules. This process regulates collective behaviors such as biofilm formation and virulence factor expression<sup>31–33</sup>. Other enriched pathways include biosynthesis of peptidoglycan<sup>34,35</sup> and various pathways related to carbohydrate metabolisms, such as starch and sucrose metabolism, galactose metabolism, and the phosphotransferase system. These pathways likely highlight strain-specific adaptations in carbon source utilization and preferences across different environments<sup>35,36</sup>.

### Distribution of paralogous genes across pan-genome groups

Paralogous genes play a fundamental role in genome evolution and the functional diversification of organisms. They provide genetic redundancy, which allows for the development of new gene functions



**Fig. 5 | Pan-genome analysis of 2794 *Streptococcus suis* strains using PGAP2.**

**a** Pie chart showing the proportion of core, shell, and cloud gene clusters; accompanying histogram illustrates the number of clusters as a function of strain frequency. **b** Rarefaction curves for the pan-genome and core genome. **c** Distribution of four quantitative parameters (Uniqueness, Minimum identity, Mean identity, and Variance) calculated by PGAP2 for all gene clusters. **d** Similarity relationships among homologous gene clusters. Unique clusters are defined as those without detectable homology to any other clusters, whereas coherent

clusters exhibit sequence similarity exceeding the predefined gene family threshold. The bar chart depicts the distribution of these two cluster types across different pan-genome categories. **e** Intra-cluster similarity among homologous genes. Conserved clusters are defined as those with sequence identity above the conservation threshold, indicating strong internal consistency. Variable clusters meet the gene family threshold but fall below the conservation threshold, indicating elevated diversity within the cluster.

through the processes of duplication and divergence<sup>37,38</sup>. To investigate this further, we introduced the concept of a paralogous ratio, which is defined as the average number of paralogous genes per strain within a homologous gene cluster. A higher paralogous coefficient indicates that these genes tend to cluster within specific strains, while a lower value suggests a more uniform distribution across all strains.

We investigated the distribution of paralogous genes in *Streptococcus suis* and found distinct patterns across different pan-genome groups (Supplementary Fig. 14A). Core gene clusters generally exhibit an even distribution of paralogous genes across various strains. In contrast, accessory gene clusters are more likely to contain multiple copies within individual strains, which reflects a statistically significant difference. Interestingly, although paralogous gene clusters are most prevalent in the cloud category, they account for only 6.75% of all cloud gene clusters. In comparison, paralogous gene clusters account for 51.54% (719 out of 1395) of all core gene clusters, with the highest proportion found in strict core clusters at 60.76% (79 out of 130)

(Supplementary Fig. 14B). These findings suggest that additional copies of core gene clusters provide essential redundancy for critical functions, thereby enhancing the survival and robustness of *S. suis* under varying conditions<sup>39</sup>.

## Discussion

In the past two decades, significant development trends have reshaped prokaryotic pan-genome research<sup>40–44</sup>. These trends highlight the growing need for advanced analytical methods and computational models to handle increasingly large datasets, understand complex genomic dynamics, and uncover the evolutionary narratives contained within the pan-genome. Although state-of-the-art methods were effective in earlier studies, they have struggled to keep up with the rapid advancements in sequencing technologies and the explosion of large-scale datasets.

PGAP2 is designed to address the challenges posed by the rapid growth in dataset size. Unlike full-graph search strategies such as



Roary<sup>22</sup> and PanTa<sup>20</sup>, PGAP2 employs a dual-level regional restriction mechanism. This approach limits the search for orthologous genes to a confined radius, greatly reducing unnecessary computations. Doing so not only accelerates the search process but also allows enough time for detailed feature analysis, thereby enhancing overall accuracy. As a result, PGAP2 can efficiently process large-scale genomic data. In our case study, PGAP2 successfully inferred the pan-genome of 2794 *Streptococcus suis* strains in just 20 minutes using 28 threads, outperforming previous methods by a significant margin. This demonstrates that large-scale pan-genome analysis is not only feasible in the high-throughput computing era but also highly efficient.

Most current methods for inferring complex gene families primarily rely on gene identity and synteny within strains, which are effective for conserved homologs like core genes. However, these criteria often fall short when dealing with more intricate gene families. Tree-based methods, such as PEPPAN<sup>17</sup> and PanX<sup>18</sup>, can analyze phylogenetic relationships to infer complex gene families. Still, they tend to be time-consuming and are unsuitable for large-scale genomic analyses. PGAP2 addresses these challenges by introducing a fine-grained feature analysis mechanism that considers strain diversity, connectivity, and bidirectional best-hit (BBH) criteria in the identification of orthologous genes. These internal features help to better constrain the inference of complex gene clusters, resulting in a more accurate and reliable process.

Moreover, PGAP2 introduces a suite of parameters that quantitatively describe the properties of homologs, allowing researchers to capture the evolutionary dynamics of gene families in a more detailed and intuitive way. This focus on quantitative analysis is essential for understanding the forces driving prokaryotic evolution, and PGAP2's ability to quantify these relationships sets it apart from other methods. Furthermore, building on the foundation of PGAP<sup>16</sup>, PGAP2 provides seamless workflows for preprocessing and post-processing (Supplementary Figs. 1–6), along with visualization tools to aid users in exploring and understanding the underlying dynamics of pan-genomes.

PGAP2 has several current limitations. First, it does not include non-coding genes in pan-genome clustering because its pipeline relies on protein-level similarity, which non-coding elements typically lack. Second, PGAP2 is sensitive to the quality of genome assemblies. It performs poorly on fragmented or incomplete assemblies, such as metagenome-assembled genomes (MAGs), where significant portions of the sequences may be missing. This lack of continuity disrupts gene order and weakens synteny signals, which are crucial for accurately distinguishing orthologs from paralogs in PGAP2. As a result, such inputs can significantly compromise the accuracy of clustering. Additionally, PGAP2 relies on both protein similarity and gene synteny among genomes. When these assumptions are violated, for example in analyses that span a genus with evolutionarily distant species, clustering accuracy may decline, and results should be interpreted cautiously (Supplementary Fig. 15). In such cases, further biological validation is recommended. Lastly, the fine-grained feature analysis used by PGAP2 is computationally intensive. While region-restricted alignment strategies help avoid unnecessary comparisons, runtime can still increase significantly for very large datasets, especially those with extensive paralogy and fragmentation. For example, analyzing the complete *E. coli* dataset (over 40,000 genomes) led to a considerable increase in both runtime and memory usage, indicating a practical scalability limit with the current configurations.

## Methods

### *Escherichia coli* strain selection for manual curation

A total of 170 complete genomes with well-defined subtype annotations were downloaded from the NCBI RefSeq database. The genome of the representative strain *Escherichia coli* K-12 substr. MG1655 was selected as the reference genome. Genome distances between the

reference genome and other genomes were calculated using Mash v2.3 with default parameters<sup>45</sup>. A phylogenetic tree based on Mash distances was constructed using Mashtree v1.2.0<sup>46</sup>, and the tree shape was adjusted with iTOL v6.9.1<sup>47</sup>, as shown in Supplementary Fig. 9. Representative strains were selected based on the genome distance gradient and their positions within the tree. In total, 50 *Escherichia coli* strains, including the reference strain K-12 substr. MG1655 and another representative strain O157:H7 str. Sakai, were selected as the initial dataset for the gold standard. Detailed information of the data curation process is provided in Fig. 3a.

All CDS sequences were extracted from genome sequences and gff annotation files by in-house scripts. All pseudogenes and CDS sequences with protein sequence lengths less than 20 were deleted. Each sequence was renamed in the format A\_B\_C to generate unique sequence identification numbers, where A represents the genome number, B represents the chromosome and plasmid number, and C represents the sequence number.

### Gene family classification of *Escherichia coli* benchmark

We curated orthologous gene families using a six-step cascade process, as outlined in Fig. 3a and Supplementary Methods. In step 1, we merged genes with identical annotations, and calculated the semantic similarity of gene products extracted from annotation files using the Python library SequenceMatcher. For gene families with an annotation similarity greater than 90%, we manually curated them to confirm if they represented the same protein. If they were determined to be identical, we combined them into the same gene family. In step 2, to address annotation differences caused by inconsistent protein names (e.g., ribonucleoside hydrolase 1 and pyrimidine-specific ribonucleoside hydrolase RihA, which are the same protein), we used BLASTp v2.15.0<sup>48</sup> to re-annotate all protein sequences against the Swiss-Prot database<sup>49</sup>, setting thresholds of identity  $\geq 70\%$ , coverage  $\geq 70\%$ , and e-value  $\leq 1e-5$ . Gene families with consistent Swiss-Prot annotations were then merged.

While homologous annotation effectively identifies types of gene families, it fails to adequately distinguish between paralogous genes because it does not take the location of genes within the genome into account. To address this, step 3 introduced a genomic context score, which simulates the degree of collinearity between genome segments by comparing the number of homologous genes within the five upstream and downstream genes surrounding each target gene. Additionally, phylogenetic trees were constructed for all assumed paralogous gene clusters using the default parameters of MAFFT v7.310<sup>50</sup> and FastTree v2.1.10<sup>51</sup>, and using ETE3 v3.1.3<sup>52</sup> used to obtain more accurate evolutionary relationships. The classification rules were as follows: If a gene was not located on the same genome as its nearest gene, paralogous genes were distinguished based on the genomic context score. In cases where the genomic context scores were consistent, paralogs were separated according to the principle of nearest distance. This process resulted in the homologous gene family classification established in step 3.

Next, we focused on unknown proteins, typically annotated as hypothetical proteins, that homologous annotations could not classify. The best matching genes from all strains were searched for these proteins rather than relying on known databases. The bidirectional best hit (BBH) of a gene was defined based on the following criteria: identity  $\geq 70\%$ , coverage  $\geq 70\%$ , and e-value  $\leq 1e-5$ . If the BBH had been correctly classified, the corresponding hypothetical protein was classified into the gene family of the BBH. If no BBH was found, the hypothetical protein was classified into a new family. In cases where the BBH was also a hypothetical protein, no classification was made at that stage. Finally, any unclassified putative proteins were grouped using MCL v14-137<sup>53</sup>, applying the same bidirectional thresholds of identity  $\geq 70\%$ , coverage  $\geq 70\%$ , and e-value  $\leq 1e-5$ . This approach ensured a clear differentiation between clusters, culminating in step 4.

Additionally, we performed a consistency check based on BBH for all classified gene clusters. This was done to verify that all BBH pairs belonged to the same gene family and to identify any proteins that may have been misclassified due to not meeting the BBH criteria. The misclassified genes were extracted and reclassified using MCL v14-137 with the same bidirectional thresholds to form step 5.

We finally addressed special genes, such as Insertion Sequences (IS) and prophage proteins, which are mobile elements that often appear randomly in genomes and do not conform to traditional homologous classification rules<sup>54,55</sup>. These genes were manually classified using a combination of genome context score, sequence similarity, and annotation information. For genes that remained ambiguous, we ensured the accuracy of our dataset by removing them, so that only correctly classified genes were retained. This process led to the creation of the final version of the gene family.

### Quality assessment of *Escherichia coli* benchmark dataset

Benchmarking the inference of orthologs typically involves several key factors, such as testing for functional consistency, inconsistencies in the species tree, conservation of gene neighborhoods, and validation against gold standard gene trees<sup>56,57</sup>. Therefore, the reliability of the benchmark was verified by the sequence similarity of proteins, the consistency of domain annotation, the average genetic distance of genes, and the association between the nearest genetic distance and the average genetic distance. First, we classified the types of gene clusters in our benchmark. We then used BLASTp to calculate the protein sequence similarity of pair-to-pair genes in each gene family and the average sequence similarity within the cluster. Next, the proteins in each gene family were annotated by InterProScan v5.63<sup>58</sup>. We calculated the semantic similarity in the same family based on the domains annotated, employing the Python package scikit-learn to compute cosine similarity. We performed alignment of nucleic acid sequences within each genome family using MAFFT v7.310<sup>59</sup>. Next, we utilized FastTree v2.1.10<sup>60</sup> to build the gene tree and employed ETE3 v3.1.3<sup>52</sup> to calculate the genetic distance between each pair of genes. Following this, we extracted representative sequences from each genome family for sequence alignment and phylogenetic tree construction. We also calculated the genetic distance between gene families and compared the average genetic distance within a gene family to that between the nearest gene family to determine if there was a significant difference.

### Characteristic analysis of *Escherichia coli* benchmark dataset

We used PanGP v1.0.1<sup>13</sup> to calculate the pan-genome size of *E. coli*, including the core genome size and the number of new genes. This was done through nonlinear fitting based on the gene presence/absence matrix of *E. coli* benchmark. To annotate all proteins and identify the functional characteristics of gene clusters, we utilized egg-nog-mapper v2.1.12 with the DIAMOND alignment algorithm, referencing the egg-NOG database v5.0.2. Next, we constructed OrgDB annotation package by R package AnnotationForge v1.40.2, and performed GO enrichment analyses for gene clusters using clusterProfiler v4.6.2<sup>61</sup> with P-value < 0.05. Finally, we visualized the COG annotations and GO terms for core genes, shell genes, and cloud genes using R scripts.

### *Salmonella enterica* benchmark dataset generation

The *Salmonella enterica* dataset was initially generated by Sean-Paul Nuccio and Andreas J. Bäumlér<sup>62</sup>. The strains were downloaded from the NCBI genome database, with accession numbers as follows: LT2: GCF\_000006945.1, SL254: GCF\_000016045.1, SL476: GCF\_000020705.1, CVM19633: GCF\_000020745.1, SL483: GCF\_000020885.1, P125109: GCF\_000009505.1, SPB7: GCF\_000018705.1, CT18: GCF\_000195995.1, Ty2: GCF\_000007545.1, ATCC-9150: GCF\_000011885.1, AKU-12601: GCF\_000026565.1, CT-

02021853: GCF\_000020925.1, 287-91: GCF\_000009525.1, SC-B67: GCF\_000008105.1, RKS4594: GCF\_000018385.1, with the corresponding database links provided in Supplementary Table 15. We generated a standardized homologous gene matrix from the supplementary table provided in their research. Only complete genes from the supplementary tables were chosen as test genes to account for differences in initial gene filtering criteria among pan-genome software tools (Supplementary Data 2).

### Representative strain and outlier selection

Identifying outliers is a critical step in PGAP2 preprocessing, achieved through two distinct homologous gene-based and ANI similarity-based screening methods.

First, Homologous Gene-Based Screening: PGAP2 uses CD-HIT<sup>63</sup> or MMseqs2<sup>64</sup> to eliminate redundancy and select representative sequences. These sequences are then aligned in an all-against-all manner to create a connectivity graph of the representative sequences. The distance between strains is represented by the strains found within their respective connected subgraphs. PGAP2 identifies all connected subgraphs that contain only a single strain and counts how many times each strain appears within these subgraphs. It calculates the third quartile (Q3) and interquartile range (IQR) of these counts, with the upper bound defined as  $Q3 + 1.5 \times IQR$ . Strains that exceed this upper bound are classified as outliers.

Second, ANI Similarity-Based Screening: Average Nucleotide Identity (ANI) values between genomes are determined using FastANI v1.1<sup>65</sup> with the default settings. To ensure reliable estimates, there must be at least 150 kb of homologous genome sequences present between the two genomes being compared. Strains with ANI values below 95% compared to a representative genome are classified as outliers. Users have the option to specify the representative genome, or it can be automatically selected by PGAP2. When PGAP2 selects a representative genome, it identifies all connected subgraphs that contain more than half of the total strains and counts how often each strain occurs within them. The strain that appears most frequently is designated as the representative genome.

The combination of strains identified by these two methods forms the final set of outliers, which is displayed by PGAP2.

### The core basis for the division of sub-clusters

PGAP2 processes each initial similarity network generated by MCL, merging nodes that are within a specified distance into sub-clusters. Each node in the sub-cluster represents a unique gene without paralogs, along with its associated members. Nodes within a sub-cluster must meet basic synteny and similarity requirements. Additionally, PGAP2 applies three additional core criteria to determine whether nodes within a subcluster should be merged:

1. Cluster Diversity: PGAP2 identifies high-quality initial core gene clusters from the initial connectivity graph. These clusters are defined as those without duplicate genes within the same strain and containing the number of strains that correspond to strict core genes. The algorithm calculates the range of similarities between nodes within these initial clusters and uses the maximum range as a threshold for merging other subclusters. Specifically, the similarity difference between any nodes in the candidate sub-clusters must not exceed this threshold.
2. Cluster Connectedness: This metric evaluates the number of edges connecting two candidate nodes for merging. Users can configure the strictness of this criterion using the ‘--sensitivity’ parameter: Strict: all child nodes between two candidate nodes must be connected, forming a fully connected subgraph; Moderate: at least 50% of the child nodes between two candidate nodes must be connected; Soft: at least one pair of child nodes must be connected.

3. Bidirectional Best Hit (BBH): If two nodes are merged into a paralogous gene cluster, the similarity between the resulting paralogous genes must exceed the maximum similarity found among the child nodes of their respective original nodes.

PGAP2 evaluates all pairs of nodes within a subcluster iteratively until all nodes are either merged or rejected. This approach ensures that the resulting clusters are accurate and biologically relevant.

### Pan-genome software selection for comparing

To assess the performance of PGAP2, we selected five popular software tools for prokaryotic pan-genome analysis: Panaroo (Version 1.5.0), PanTa (Version 1.0), PEPPAN (Version 1.0.5), PPanGGOLiN (Version 2.1.0), and Roary (Version 3.12.0). These tools were chosen based on their popularity in the field, their distinct methodological principles, with Roary and PanTa relying on conserved gene neighborhood approaches, Panaroo and PPanGGOLiN employing graph-based models, and PEPPAN applying a tree-based phylogenetic framework, as well as their publication dates, which highlight their relevance and influence in pan-genome research.

### Performance evaluation using simulated data

All test datasets were generated using SimPan. The parameter `--idenOrtholog` was varied, decreasing from 0.99 to 0.9 in steps of 0.02 to simulate different orthologous thresholds. Meanwhile, the `--idenParalog` parameter was reduced from 0.9 to 0.6 in steps of 0.1 to simulate varying paralogous thresholds. All other parameters were set to their default values. The default parameters were utilized for all pan-genome software tools tested, with Panaroo operating in “sensitive” mode and PGAP2 with “--ins” parameters to turn off insertion sequence judgment.

### Performance evaluation using gold standard data

When evaluating software performance using the *E. coli* gold-standard dataset, the paralogous threshold was set to 0.7 for all software tools that incorporated this parameter. For the *S. enterica* gold-standard dataset, the paralogous threshold was set to 0.9. Additionally, since both gold-standard datasets offered a “fully split paralogous gene clusters” mode as the reference matrix, PGAP2 was configured with the `--exhaust_ortho` option to ensure that no in-paralogs were included in the output.

### Reporting summary

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

### Data availability

All genome sequences used in this study were obtained from the NCBI RefSeq database (<https://www.ncbi.nlm.nih.gov/refseq/>) and are publicly available. The accessions of all datasets analyzed in this study are provided in Supplementary Table 9 (for *Escherichia coli*) and Supplementary Table 15 (for *Salmonella*). The curated benchmark datasets generated in this study are provided in Supplementary Data 1 and Supplementary Data 2. The scripts used for constructing the benchmark and performing evaluations are available at GitHub ([https://github.com/bucongfan/PGAP2\\_benchmark](https://github.com/bucongfan/PGAP2_benchmark)). The curated GFF3 files and the final PAV matrices are archived in Figshare (<https://doi.org/10.6084/m9.figshare.29908154.v1>).

### Code availability

The source code of PGAP2 is written in Python and available at BioCode (<https://ngdc.cncb.ac.cn/biocode/tool/BT7758>) and GitHub (<https://github.com/bucongfan/PGAP2>) under the MIT license. The source code with full software environment used in this study is available in Code Ocean at <https://doi.org/10.24433/CO.9288245.v2>.

## References

1. Arnold, B. J., Huang, I. T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nat. Rev. Microbiol.* **20**, 206–218 (2022).
2. Power, J. J. et al. Adaptive evolution of hybrid bacteria by horizontal gene transfer. *Proc. Natl Acad. Sci. USA* **118**, e2007873118 (2021).
3. Mohite, O. S., Lloyd, C. J., Monk, J. M., Weber, T. & Palsson, B. O. Pangenome analysis of Enterobacteria reveals richness of secondary metabolite gene clusters and their associated gene sets. *Synth. Syst. Biotechnol.* **7**, 900–910 (2022).
4. Park, S. C., Lee, K., Kim, Y. O., Won, S. & Chun, J. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Front Microbiol.* **10**, 834 (2019).
5. Decano, A. G. & Downing, T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci. Rep.* **9**, 17394 (2019).
6. Belbahri, L. et al. Comparative genomics of *Bacillus amyloliquefaciens* strains reveals a core genome with traits for habitat adaptation and a secondary metabolites rich accessory genome. *Front Microbiol.* **8**, 1438 (2017).
7. Sigalova, O. M. et al. Chlamydia pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction. *BMC Genomics* **20**, 710 (2019).
8. Saw, J. H. W. et al. Pangenomics analysis reveals diversification of enzyme families and niche specialization in globally abundant SAR202 bacteria. *mBio* **11**, e02975–02919 (2020).
9. Mageiros, L. et al. Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. *Nat. Commun.* **12**, 765 (2021).
10. Thorpe, H. A., Bayliss, S. C., Sheppard, S. K. & Feil, E. J. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience* **7**, 1–11 (2018).
11. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).
12. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* **47**, 5539–5549 (2019).
13. Zhao, Y. et al. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics* **30**, 1297–1299 (2014).
14. Hernandez-Plaza, A. et al. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).
15. Galperin, M. Y. et al. COG database update 2024. *Nucleic Acids Res.* **53**, D356–D363 (2024).
16. Zhao, Y. et al. PGAP: pan-genomes analysis pipeline. *Bioinformatics* **28**, 416–418 (2012).
17. Zhou, Z., Charlesworth, J. & Achtman, M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.* **30**, 1667–1679 (2020).
18. Ding, W., Baumdicker, F. & Neher, R. A. panX: pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
19. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
20. Le, D. Q. et al. Efficient inference of large prokaryotic pangenomes with PanTA. *Genome Biol.* **25**, 209 (2024).
21. Gautreau, G. et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.* **16**, e1007732 (2020).
22. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
23. Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K. & Feil, E. J. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* **8**, giz119 (2019).



24. Tantoso, E. et al. To kill or to be killed: pangenome analysis of *Escherichia coli* strains reveals a tailocin specific for pandemic ST131. *BMC Biol.* **20**, 146 (2022).
25. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced *Escherichia coli* genomes. *Micro. Ecol.* **60**, 708–720 (2010).
26. Yang, T. & Gao, F. High-quality pan-genome of *Escherichia coli* generated by excluding confounding and highly similar strains reveals an association between unique gene clusters and genomic islands. *Brief. Bioinform* **23**, bbac283 (2022).
27. Lee, I. P. A. & Andam, C. P. Frequencies and characteristics of genome-wide recombination in *Streptococcus agalactiae*, *Streptococcus pyogenes*, and *Streptococcus suis*. *Sci. Rep.* **12**, 1515 (2022).
28. Zhang, A. et al. Comparative genomic analysis of *Streptococcus suis* reveals significant genomic diversity among different serotypes. *BMC Genomics* **12**, 523 (2011).
29. Xia, Y. et al. Isolation, identification, genomic diversity, and antimicrobial resistance analysis of *Streptococcus suis* in hubei province of China from 2021 to 2023. *Microorganisms* **12**, 917 (2024).
30. Guo, G. et al. Pan-genome analysis of *Streptococcus suis* serotype 2 revealed genomic diversity among strains of different virulence. *Transbound. Emerg. Dis.* **68**, 637–647 (2021).
31. Prescott, R. D. & Decho, A. W. Flexibility and Adaptability of Quorum Sensing in Nature. *Trends Microbiol* **28**, 436–444 (2020).
32. Moreno-Gamez, S., Hochberg, M. E. & van Doorn, G. S. Quorum sensing as a mechanism to harness the wisdom of the crowds. *Nat. Commun.* **14**, 3415 (2023).
33. Williams, P. Quorum sensing, communication and cross-kingdom signalling in the bacterial world. *Microbiol. (Read.)* **153**, 3923–3938 (2007).
34. Garde, S., Chodiseti, P. K. & Reddy, M. Peptidoglycan: Structure, Synthesis, and Regulation. *EcoSal* **9**, eESP-0010–eESP-2020 (2021).
35. Mueller, E. A. & Levin, P. A. Bacterial Cell Wall Quality Control during Environmental Stress. *mBio* **11**, e02456–02420 (2020).
36. Armitage, J. P. Behavioral responses in bacteria. *Annu Rev. Physiol.* **54**, 683–714 (1992).
37. Kleinjan, D. A. et al. Subfunctionalization of duplicated zebrafish *pax6* genes by cis-regulatory divergence. *PLoS Genet* **4**, e29 (2008).
38. Conrad, B. & Antonarakis, S. E. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev. Genomics Hum. Genet* **8**, 17–35 (2007).
39. Gevers, D., Vandepoele, K., Simillon, C. & Van de Peer, Y. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* **12**, 148–154 (2004).
40. Kalan, L. R. et al. Strain- and Species-Level Variation in the Microbiome of Diabetic Wounds Is Associated with Clinical Outcomes and Therapeutic Efficacy. *Cell Host Microbe* **25**, 641–655 e645 (2019).
41. Thorpe, H. A. et al. A large-scale genomic snapshot of *Klebsiella* spp. isolates in Northern Italy reveals limited transmission between clinical and non-clinical settings. *Nat. Microbiol* **7**, 2054–2067 (2022).
42. Smith, A. B. et al. Enterococci enhance *Clostridioides difficile* pathogenesis. *Nature* **611**, 780–786 (2022).
43. Sands, K. et al. Characterization of antimicrobial-resistant Gram-negative bacteria that cause neonatal sepsis in seven low- and middle-income countries. *Nat. Microbiol* **6**, 512–523 (2021).
44. Rida, T. et al. Pan-Genome Analysis of Oral Bacterial Pathogens to Predict a Potential Novel Multi-Epitopes Vaccine Candidate. *Int J. Environ. Res Public Health* **19**, 8408 (2022).
45. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
46. Katz, L. S. et al. Mashtree: a rapid comparison of whole genome sequence files. *J. Open Source Softw.* **4**, <https://doi.org/10.21105/joss.01762> (2019).
47. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res* **52**, W78–W82 (2024).
48. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
49. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–d489 (2021).
50. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
51. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).
52. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
53. Van Dongen, S. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).
54. Siguier, P., Gournayre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
55. Rousset, F. et al. Phages and their satellites encode hotspots of antiviral systems. *Cell host microbe* **30**, 740–753.e745 (2022).
56. Altenhoff, A. M. & Dessimoz, C. Inferring orthology and paralogy. *Methods Mol. Biol. (Clifton, NJ)* **855**, 259–279 (2012).
57. Nevers, Y. et al. The Quest for Orthologs orthology benchmark service in 2022. *Nucleic acids Res.* **50**, W623–w632 (2022).
58. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic acids Res.* **51**, D418–d427 (2023).
59. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
60. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
61. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov. (Camb. (Mass))* **2**, 100141 (2021).
62. Nuccio, S. P. & Baumber, A. J. Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *mBio* **5**, e00929–00914 (2014).
63. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
64. Mirdita, M., Steinegger, M. & Soding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **35**, 2856–2858 (2019).
65. Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

## Acknowledgements

This work was supported by grants from the National Key Research Program of China [2023YFC2604402 to C.B. and 2020YFA0907001 to J.X.]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDA0460200 to J.X.]; Natural Science Foundation of China [32170669 to J.X. and 32500070 to C.B.]. Funding for open access charge: National Key Research Program of China [2023YFC2604402 to C.B.]. We thank Dr. Xinchang Zheng for their valuable discussions and insightful suggestions.



## Author contributions

C.B.: software, methodology, manuscript writing; H.Z.: data curation, manuscript writing; F.Z., W.L., and Q.L.: algorithm optimization; H.G., J.Z., R.X., F.L., and Z.Z.: revising; J.X., and Z.J.: conceptualization, revising, funding acquisition, and supervision. All authors have read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-64846-5>.

**Correspondence** and requests for materials should be addressed to Zhong Jin or Jingfa Xiao.

**Peer review information** *Nature Communications* thanks Guillaume Gautreau, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025