nature microbiology

Resource

https://doi.org/10.1038/s41564-024-01723-9

Single-cell transcriptomics across 2,534 microbial species reveals functional heterogeneity in the rumen microbiome

Received: 8 October 2023

Accepted: 7 May 2024

Published online: 12 June 2024

Check for updates

Minghui Jia[1,2,3,9], Senlin Zhu[1,2,3,9], Ming-Yuan Xue[1,2,8,9], Hongyi Chen[1,2], Jinghong Xu[1,2], Mengdi Song[4,5,6], Yifan Tang[1,2], Xiaohan Liu[1,2], Ye Tao[7], Tianyu Zhang[4,5,6], Jian-Xin Liu[1,2], Yongcheng Wang[4,5] ✉ & Hui-Zeng Sun[1,2,3] ✉

Deciphering the activity of individual microbes within complex communities and environments remains a challenge. Here we describe the development of microbiome single-cell transcriptomics using droplet-based single-cell RNA sequencing and pangenome-based computational analysis to characterize the functional heterogeneity of the rumen microbiome. We generated a microbial genome database (the Bovine Gastro Microbial Genome Map) as a functional reference map for the construction of a single-cell transcriptomic atlas of the rumen microbiome. The atlas includes 174,531 microbial cells and 2,534 species, of which 172 are core active species grouped into 12 functional clusters. We detected single-cell-level functional roles, including a key role for *Basfia succiniciproducens* in the carbohydrate metabolic niche of the rumen microbiome. Furthermore, we explored functional heterogeneity and reveal metabolic niche trajectories driven by biofilm formation pathway genes within *B. succiniciproducens*. Our results provide a resource for studying the rumen microbiome and illustrate the diverse functions of individual microbial cells that drive their ecological niche stability or adaptation within the ecosystem.

Over the years, microbiome research has achieved tremendous advancements driven by culture-independent meta-omics approaches[1–3]. Metagenomic binning techniques have been a milestone in the exploration of unculturable microbial genomes (that is, metagenome-assembled genomes (MAGs))[4,5], allowing for a deeper understanding of the functions of complex microbial environments, such as the human gut[6,7], rumen[8] and ocean[9]. However, the discovery of functional redundancy[10] and microbial heterogeneity[11] are major challenges in obtaining groundbreaking insights. New approaches are needed to address the issues of resolution (single-cell functional heterogeneity), effectiveness (RNA functionality) and accuracy (high throughput) in microbial studies, rendering them crucial for the next era of microbial research.

Single-cell microbiological techniques have emerged as potential solutions. For instance, single-cell genomic approaches have facilitated the identification of genomic information for revealing microbial

[1]Institute of Dairy Science, College of Animal Sciences, Zhejiang University, Hangzhou, China. [2]Key Laboratory of Molecular Animal Nutrition, Ministry of Education, Zhejiang University, Hangzhou, China. [3]Key Laboratory of Dairy Cow Genetic Improvement and Milk Quality Research of Zhejiang Province, Zhejiang University, Hangzhou, China. [4]Liangzhu Laboratory, Zhejiang University, Hangzhou, China. [5]Department of Laboratory Medicine, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. [6]M20 Genomics, Hangzhou, China. [7]Shanghai Biozeron Biotechnology Company, Shanghai, China. [8]Present address: Xianghu Laboratory, Hangzhou, China. [9]These authors contributed equally: Minghui Jia, Senlin Zhu, Ming-Yuan Xue. ✉e-mail: yongcheng@zju.edu.cn; huizeng@zju.edu.cn

members[12]. Nevertheless, they are not able to identify active functions and expressed genes[13]. Detecting individual gene expression patterns using single-cell RNA sequencing (scRNA-seq) is essential for elucidating the mechanisms underlying such heterogeneity in microbial cells[14]. Many efforts have been made in microbial scRNA-seq, including prokaryotic expression profiling by tagging RNA in situ and sequencing (PETRI-seq)[14], microbial split-pool ligation transcriptomics (microSPLiT)[15], eukaryotic bacterial droplet-based scRNA-seq (BacDrop)[16] and droplet-based high-throughput single-microbe RNA-seq (smRandom-seq)[17]. However, these methods focus primarily on cellular heterogeneity in simple synthetic microbial communities with known members. There remains a huge technical and knowledge gap in exploring the active functional roles of specific organisms in complex microbial environments with a tremendous number of unknown and unculturable members. Therefore, better tools for microbiome scRNA-seq that can be applied to uncultivable or under-characterized microbial ecosystems are urgently needed.

The rumen microbiome—one of the most complex and under-investigated microbial habitats—is responsible for degrading inedible plant biomass to produce high-quality protein products (meat and milk) while generating considerable environmental problems[18]. The rumen microbiome represents a complex environment with limited pangenome information and transcriptional data. Due to its complex taxonomy, large functional redundancy and strict anaerobic nature, the current understanding of the rumen microbiome is still limited and is restricted to the Hungate1000 Project[19] and several metagenomic binning studies[8,20]. In this Resource, using the rumen microbiome as an ideal model, we create a rumen microbial pangenome reference (the Bovine Gastro Microbial Genome Map (BGMGM)) by employing rumen metagenomic sequencing of dairy cows and collecting publicly available cultured rumen microbial genomes and MAGs. By integrating random primer-based droplet scRNA-seq and BGMGM-based computational analysis, we develop microbiome single-cell transcriptomics (MscT) to reveal the single-cell functionalities of rumen microbiota. This study will hold innovative significance not only for microbiological research techniques but also for addressing global issues such as ecological dynamics, enzyme resource exploration and large-scale industrial production of lignocellulosic biofuels.

Results

Reference pangenomes of the rumen microbiome

We constructed a bovine gastrointestinal microbial genome database from public resources[8,19–27] (see Methods) and newly sequenced samples (Fig. 1 and Extended Data Fig. 1) and named it the BGMGM. The BGMGM comprises 2,311 animal samples covering ten different gastrointestinal segments, with the rumen being the most dominant (more than 1,480 samples and 29,225 genomes) (Supplementary Table 1). This map contains 47,241 microbial genomes (Supplementary Tables 2 and 3), including 410 cultured genomes and 46,831 MAGs, which successfully met strict quality control criteria (see Methods). After genome dereplication with a 95% average nucleotide identity cutoff, a total of 13,572 non-redundant genomes were retained, including 5,676 high-quality genomes and 7,896 medium-quality genomes. High-quality genomes displayed significantly higher N50 values and fewer scaffolds than medium-quality genomes, with a quality score (QS) of ≥75 ($P < 0.001$ for N50 and $P < 0.001$ for scaffolds; Fig. 2a), supporting sustained results even when compared with the relatively high-quality parts of medium-quality genomes.

After utilizing the Genome Taxonomy Database Toolkit[28] (GTDB-Tk version 2.3.2) for species annotation, 7,545 (55.6%) genomes were identified as known species. Notably, most of these genomes were predominantly from the phylum Bacillota (4,181), followed by Bacteroidota (1,916) and Pseudomonadota (328) (Supplementary Table 4). Compared with large-scale and well-established gut microbial pangenomes in humans[6,7], bovine gastrointestinal studies are still scarce. The to-date largest holistic collection—de-replicated microbial genomes identified by Watson[29]—was retrieved from 33,813 public rumen MAGs. Our BGMGM obtained an 80% increase in de-replicated putative species-level genomes (13,572 versus 7,533), among which high-quality genomes were elevated by 111% from 2,696 to 5,676. The high proportion of high-quality genomes in the BGMGM guarantees further use in single-cell transcriptomics research.

To expand our understanding of rumen microbial functions, we constructed a catalogue of protein-coding genes from the BGMGM. Collectively, 25,898,014 genes were predicted from 13,572 non-redundant genomes. After gene dereplication with a 95% average nucleotide identity cutoff, 23,755,235 genes (91.7%) were retained. We discovered 18,046,712 (76% of the total non-redundant genes) functional genes annotated by at least one database (Fig. 2b and Extended Data Fig. 2a). Functional annotation in the Clusters of Orthologous Genes (COG) database spans 22 functional categories, with the classifications of carbohydrate transport and metabolism, cell wall/membrane/envelope biogenesis and translation being the most enriched (Fig. 2c). In total, 3,112 (22.9%) near-complete genomes (including 3,065 Bacteria genomes and 47 Archaea genomes) were selected with high completeness (mean ± s.d. = 96.10 ± 2.49%) and low contamination (mean ± s.d. = 0.43 ± 0.47%). We visualized them in Fig. 2d and Extended Data Fig. 2b, respectively. The high-quality genome set, large number of annotated genes and high gene annotation rate provided a solid basis for microbial pangenome mapping and functional investigation of single-cell transcriptomic data.

The rumen microbiome single-cell functional landscape

To annotate the microbiome transcriptomics of the complex rumen microbial ecosystem using scRNA-seq, we developed a strategy for microbial pangenome mapping and functional cluster identification (Extended Data Fig. 3a). We captured more than 200,000 rumen microbial cells using our previously developed droplet-based single-microbe RNA-seq method[17] with optimized random primers and a microfluidic barcoding platform. After relatively strict quality control (see Methods), 174,531 high-quality cells were retained, with a median number of 4,611 unique molecular identifiers and 182 unique genes per cell (Extended Data Fig. 3b). The large cell numbers and unique gene numbers (288,268) ensure the applicability of MscT to microbiome investigation. After normalization and a series of benchmarking (Extended Data Fig. 4a), we performed clustering analysis with batch effect correction on all 174,531 cells and identified 12 functional clusters (Fig. 3a). These functional clusters were annotated based on the biological functions of the specifically expressed genes (Supplementary Table 5). For example, cells in the HSP90^{+} high metabolic activity cell (HMAC) functional cluster specifically expressed the *CowSGB-6222-c11-2* gene, which encodes the HSP90 protein associated with ATP utilization, suggesting that this functional cluster possesses metabolic activity. Meanwhile, the proportion of metabolically active genes in the cells of this cluster is relatively high compared with that of other cells (Supplementary Table 6); therefore, we named this functional cluster HSP90^{+} HMACs (other functional clusters were named similarly, as detailed in Supplementary Tables 5 and 6). The 12 functional clusters represent a classification of rumen microbes at the single-cell RNA level, with high overlap between different samples (Extended Data Fig. 4b), demonstrating the robustness and reproducibility of the functional cluster analysis strategy. Due to redundancy in most microbial functions[30], taxonomic analysis is not sufficient to describe the functional heterogeneity that exists[31]. Therefore, a functional group-centred approach to microbial ecology research has been proposed[32]. According to current research, the functional groups exhibit stability[33], dynamic equilibrium[34] and complex interactions[35], similar to the functional clusters we identified.

Unlike other microbial scRNA-seq studies that explore communities with known members[14–16], our study investigated previously unknown species in a complex community. MscT identified 2,534

https://doi.org/10.1038/s41564-024-01723-9

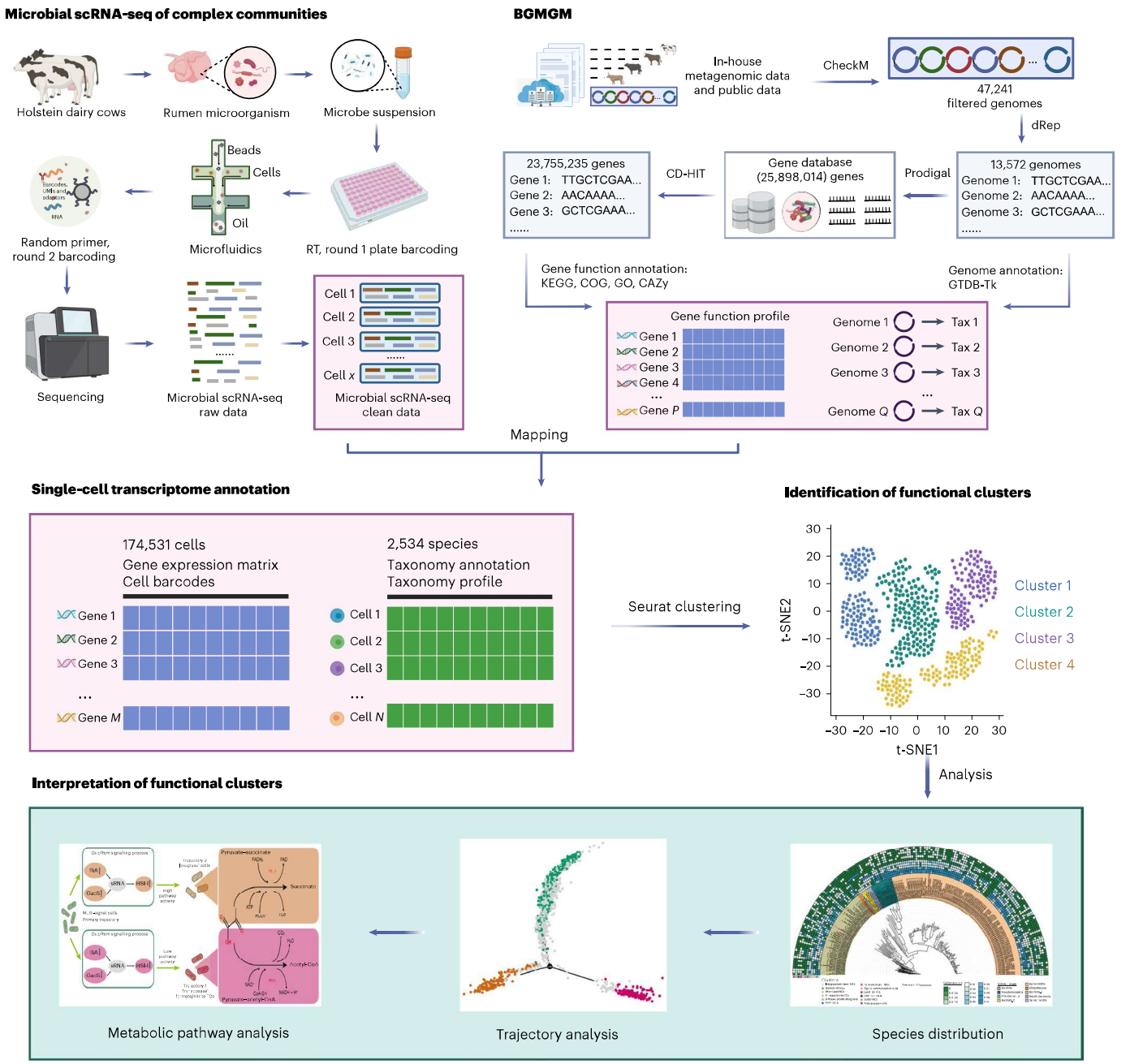


Fig. 1 | Overall workflow of MscT. This pipeline includes the microbial scRNA-seq of complex communities, BGMGM construction, transcriptomic annotation at the single-cell level and identification and interpretation of functional clusters. RT, Reverse transcription; GO, Gene Ontology; sRNA, small RNA; TCs, transporter cells; t-SNE, t-distributed stochastic neighbour embedding; UMIs, unique molecular identifiers. Figure created with BioRender.com.

species (1,849 were known). High species diversity allows for a more comprehensive understanding of rumen microbial activity. We defined species that contained more than 100 cells as core active microbial species (Fig. 3b) and explored the distribution of 172 such species in the 12 functional clusters. We found that 164 core active microbial species were distributed in more than one functional cluster and 38 core active microbial species existed in any one functional cluster. Certain species were preferentially involved in a particular functional cluster. For example, the cells of *Desulfovibrio sp016284885* were predominantly identified as sulfur metabolic cells, whereas the cells of *Sodaliphilus sp900318205* were predominantly identified as replication protein A-positive (RPA^+) lipid metabolic cells (Fig. 3c). The distribution patterns demonstrated stable performance in classifying different microbial species.

To summarize, here we benchmarked MscT in a complex microbial community and identified 12 functional clusters with distinct transcriptomic patterns, which advances our understanding of uncultivable microbial ecosystems and offers a high-resolution approach to exploring the active functions of unknown microbial environments systematically and holistically.

Heterogeneity and interaction of functional clusters

After identifying the 12 functional clusters from the MscT data, we further characterized the heterogeneity of the biological pathways in which their marker genes were involved. We present the biological processes/structures of 23 marker genes, as characterized by co-upregulation of genes involved in: (1) carbohydrate transport and metabolism; (2) replication, recombination and repair; (3) peptide

https://doi.org/10.1038/s41564-024-01723-9

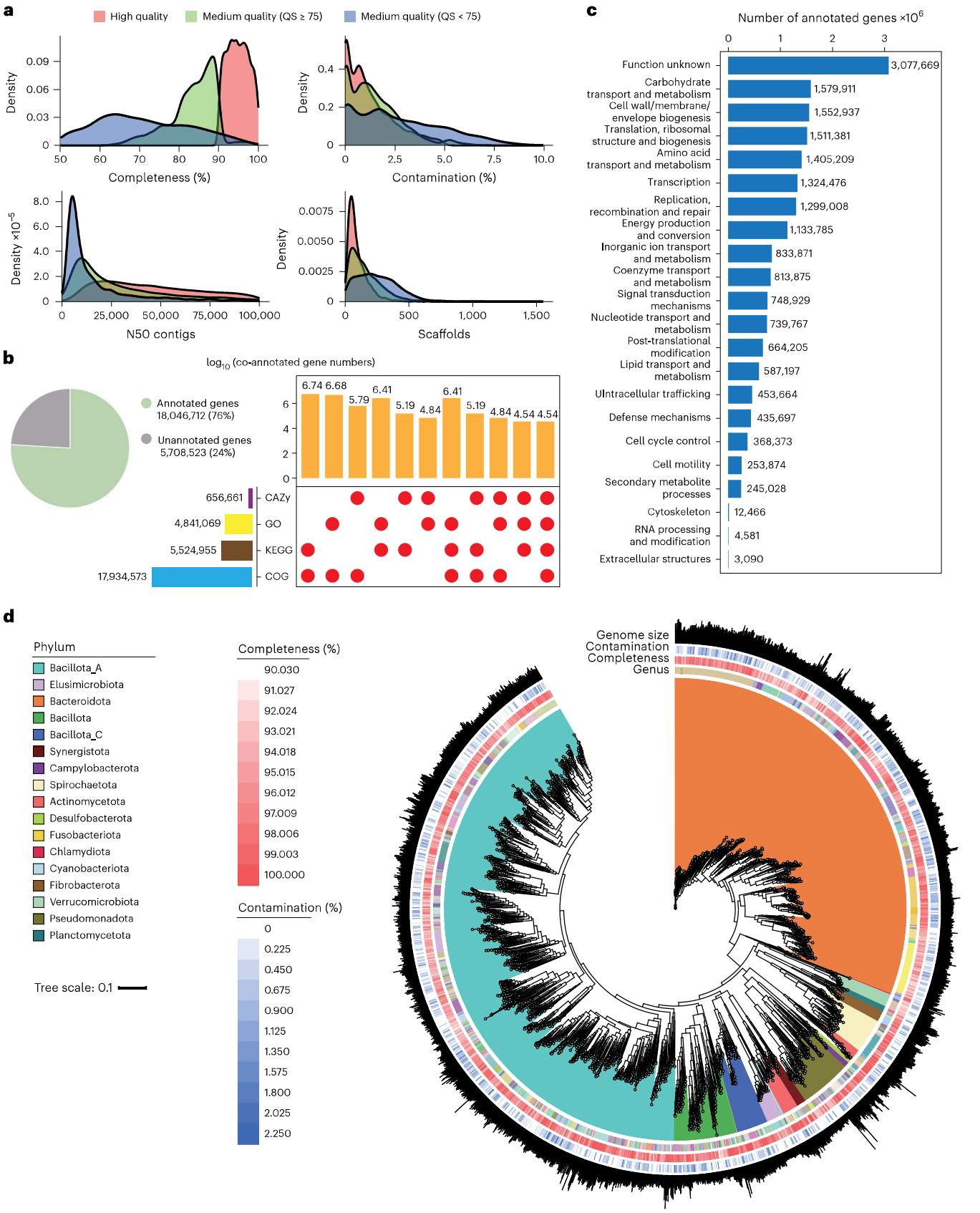


Fig. 2 | Reference pangenomes of the rumen microbiome with 13,572 non-redundant genomes (average nucleotide identity > 95%). a, Distribution of genome quality in the MAGs database. **b**, Annotation results for functional genes from the BGMGM. The COG, KEGG, Gene Ontology and CAZy databases were used. The vertical bars and dot plot present the number of functional genes annotated by different combinations of databases. The horizontal bars to the left present the total number of functional genes annotated by each database. **c**, COG pathway annotations of functional genes from the BGMGM. **d**, Phylogenetic tree of 3,065 near-complete (completeness > 90%; contamination < 5%; quality score > 100) bacterial genomes.

https://doi.org/10.1038/s41564-024-01723-9

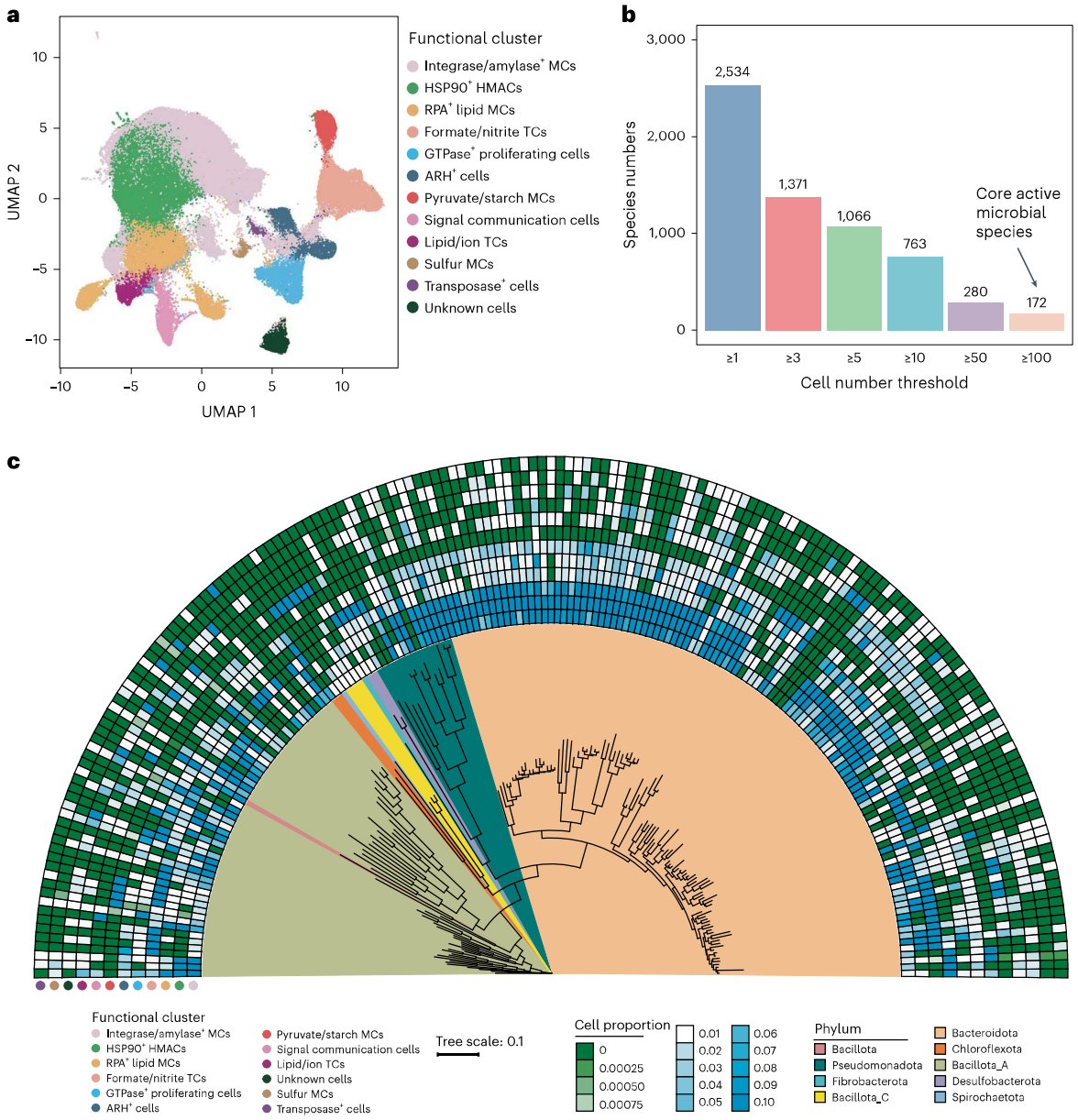


Fig. 3 | Single-cell functional landscape of the rumen microbiome. a, The uniform manifold approximation and projection (UMAP) plot for 12 functional clusters identified from a total of 174,531 cells and 288,268 genes. **b**, Species numbers of different cell number thresholds. **c**, Distribution of functional clusters and a phylogenetic tree of 172 highly cell-abundant species in 12 functional clusters. The outer blue/green semicircles display the proportion of cells in each functional cluster for each species. MCs, metabolic cells.

transportation; (4) signal transduction and metabolism; (5) lipid transport and metabolism; (6) transcription; (7) formate/nitrite transportation; (8) proliferation, stress response and ribosome biogenesis; (9) modification, protein turnover and chaperones; (10) energy and sulfur metabolism; (11) signal transduction mechanisms; (12) SecY translocase; and (13) inorganic ion transportation (Fig. 4a and Supplementary Table 5). Based on the annotated marker genes for each functional cluster, we present the specific biological pathways (Fig. 4b). For instance, sulfur metabolic cells specifically expressed the genes *CowSGB-4309-c57-3* and *CowSGB-4309-c116-5*, which encode two important proteins (AprA and DsrA) involved in the sulfur metabolic biological pathway. Therefore, we named this cluster sulfur metabolic cells and present the sulfur metabolic biological pathway in Fig. 4b. We further found that species composition varied greatly within the same cluster, reflecting the fact that different species may perform similar active functions. Meanwhile, cells from the same species were distributed in different functional clusters, suggesting that the cellular activity differed individually within species (Fig. 4a and Extended Data Fig. 4c). Our results suggest that the functional heterogeneity of microbial functional clusters stems from changes in gene expression and may be determined by differences in species composition as well as individual cell activities.

Based on the above finding, we further explored the functional heterogeneity of the same species. We extracted the cell clusters with more than 5,000 cells and more than 500 species (that is, integrase/amylase$^+$ metabolic cells, HSP90$^+$ HMACs, RPA$^+$ lipid metabolic cells, formate/nitrite transporter cells, GTPase$^+$ proliferating cells and autosomal recessive ADP-ribosylglycohydrolase-positive (ARH$^+$) cells; Extended Data Fig. 5a). Within these six cell clusters, we extracted 89 species that were distributed in more than three cell clusters with a minimum of ten cells per cluster. We used the functional gene proportion (FGP) in a single cell to determine the functional activity of each cell for a certain

https://doi.org/10.1038/s41564-024-01723-9

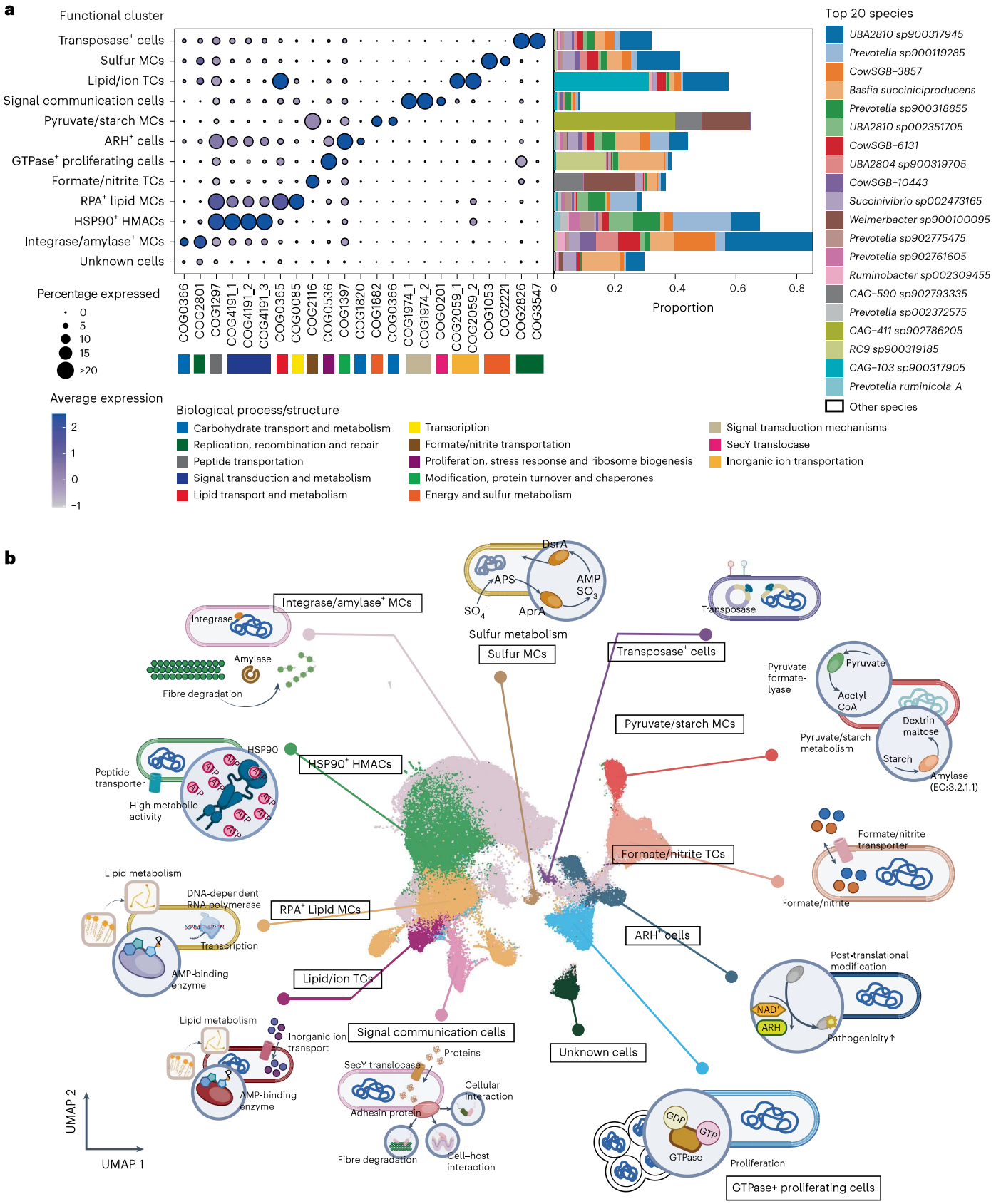


Fig. 4 | Pathway heterogeneity of functional clusters. a, Marker genes and species abundance of 12 functional clusters. **b**, Schematic of the biological processes/structures of the 12 functional groups. Panel **b** created with BioRender.com.

pathway. We calculated the FGPs for each COG pathway and analysed the inter-cluster differences of the same species. As an example, we visualized the *P* values of eight species in a heatmap (Extended Data Fig. 5a; *P* values for all species are detailed in Supplementary Table 7) and found that the inter-cluster differences varied across species, suggesting the presence of key species whose functional roles changed considerably in microbial ecosystems. We found that there were 17 COG pathways in *Basfia succiniciproducens* displaying significantly different FGPs between clusters ($P < 0.0001$). As an example, we performed inter-cluster multiple comparisons of FGPs for three important metabolic pathways in *B. succiniciproducens* (Extended Data Fig. 5b–d). The results suggest that cells of *B. succiniciproducens* in different functional clusters enact distinct roles. Next, we combined both cluster and species information into certain units (named as 'cluster—species', such as HSP90$^+$ HMACs—*B. succiniciproducens*) and performed cellular interaction network analysis on them. Within all cells, we extracted 213 units that were distributed in more than three samples with a minimum of ten cells per sample. We found a total of 519 interactions among these units (Extended Data Fig. 6). Notably, interactions were found between HSP90$^+$ HMACs—*B. succiniciproducens* and six other units, including units from the same or different clusters and species, suggesting broad ecological associations between clusters and between species. The deeper analyses of *B. succiniciproducens* are shown in Figs. 5 and 6. Overall, our results reveal differences in gene expression, function and species composition across clusters, as well as the interactions between clusters and between species, which are key to understanding the characterization of microbial functional clusters.

Single-cell metabolic insights of the rumen microbiome

After discovering the heterogeneity and interactions between different functional clusters, we further explored the heterogeneity between cells in the same functional cluster. Due to the importance of rumen microbial fermentation and the vital role of carbohydrate metabolism in this process, a large number (50,199) of HMACs were selected. HMACs showed higher overall metabolic FGPs (0.356 versus 0.048; $P < 0.01$) and carbohydrate metabolic FGPs (0.100 versus 0.033; $P < 0.01$) than the other cells (Fig. 5a). We used the gene expression matrix of the HMACs for re-clustering analysis after the more refined normalization and benchmarking processes. Based on the specifically expressed genes, ten sub-functional clusters were identified within the HMACs: helix-turn-helix-positive (HTH$^+$) HMACs, peptide transporter HMACs, integrase$^+$ HMACs, motility HMACs, membrane protein$^+$ HMACs, lipid metabolism HMACs, secretion HMACs, TonB-linked protein$^+$ HMACs, multi-signal HMACs and His kinase A$^+$ HMACs (Fig. 5b,c and Supplementary Table 5). To further explore the active roles of these HMACs in carbohydrate metabolism during rumen microbial fermentation, we refined the classic carbohydrate metabolic pathways[36], from fibre substrate (pectin, cellulose, glucan, mannan and xylan) to volatile fatty acid production (acetate, propionate and butyrate), by connecting the key intermediate metabolite pyruvate as the core[37] (Fig. 5d). We extracted accurately identified cells (cells with accurate species annotation, totalling 5,636 cells; see Methods) from the ten sub-functional clusters for subsequent analysis (Supplementary Table 8). Among the ten sub-functional clusters, we observed distinct heterogeneity in the carbohydrate metabolic pathways (Extended Data Figs. 7 and 8). According to the average FGPs of the ten sub-functional clusters, the cluster His kinase A$^+$ HMACs showed the highest metabolic activity and largest proportion of active cells in the following two processes: (1) from cellulose to glucose; and (2) from glucan to glucose. This was despite it consisting of a small number of cells (0.97%; 55/5,636 cells). The heterogeneity between sub-functional clusters reflects the re-divisibility of functional clusters, indicating that the resolution of cell types can be gradually improved by re-clustering analysis. From the analysis, we revealed that HMACs in the rumen (a certain sub-population of cells exerting patterns of high metabolism) play an important role in the classic carbohydrate metabolic pathways, from those involved with fibre content to those involved with volatile fatty acids, which renews our knowledge of rumen metabolic functions.

The production of propionate from pectin is an important pathway of fibre degradation in the rumen and was found to be the limiting pathway in crop by-product utilization in our previous study[38]. We selected four clusters ((1) integrase$^+$ HMACs; (2) HTH$^+$ HMACs; (3) peptide transporter HMACs; and (4) motility HMACs) with the highest numbers of cells (1,988, 1,462, 1,108 and 265 cells, respectively) and calculated their FGPs in four respective continuous steps: (1) pectin metabolized to produce pyruvate; (2) pyruvate metabolized to produce succinate; (3) succinate metabolized to produce propionyl coenzyme A; and (4) coenzyme A metabolized to produce propionate (Fig. 5e). Consistent with previous results, the change in pectin metabolic FGPs for these four cell clusters was distinctly heterogeneous (Fig. 5e). Because functional heterogeneity was detected between clusters of *B. succiniciproducens* in Extended Data Fig. 5, we further investigated whether the functions of *B. succiniciproducens* were consistent with the clusters to which it belonged. We found that the change in average FGPs of *B. succiniciproducens* was similar to that of the integrase$^+$ HMAC cluster, with a peak occurring during the conversion of pyruvate to succinate (Fig. 5e). The current research suggests that microbial functional group succession results from metabolically induced habitat changes, specifically in terms of the increased or decreased abundance of functional groups. The species varied consistently with functional clusters, which indicates the potential transformation of *B. succiniciproducens* cells from one to another functional group, which is further explored in Fig. 6.

Cellular functional trajectories of *B. succiniciproducens*

B. succiniciproducens is a major producer of succinic acid using glucose as a substrate[39]. As a core member of the rumen microbiome[40], *B. succiniciproducens* was detected in all of the samples of this study and found to interact with *Prevotella* species, which were considered to be keystone microbes exerting similar roles. Although the important functions of *B. succiniciproducens* have been recognized, its metabolic patterns in ruminal microbial communities are under-characterized. The high-quality genome of *B. succiniciproducens* (completeness = 100%; contamination = 0) in the BGMGM ensures the accurate use of MscT. Using MscT, 5,591 cells were accurately annotated as *B. succiniciproducens*. We extracted single-cell transcriptomics of these cells for clustering and identified eight functional clusters (Fig. 6a and Extended Data Fig. 9a,b), indicating that microbes at the species level can be further categorized into different types. Since microbial cells also have a continuous progression of biological processes based on gene expression, we performed pseudo-time analysis[41,42] and found that the functional clusters were explicitly distributed on three different trajectories of *B. succiniciproducens* cells (Fig. 6b). We extracted the clusters with the highest number of cells on each trajectory for further analysis: transposase$^+$ formate/nitrite transporter cells (trajectory1; 131 cells), integrase$^+$ cells (trajectory 2; 206 cells) and multi-signal cells (primary trajectory; 1,316 cells). These three clusters were annotated by their specifically expressed genes, such as the transposase domain gene[43], formate/nitrite transporter family gene[44] and integrase domain gene[45] (details in Supplementary Table 5). Based on the biological processes of the rumen microbiome and the dynamics from substrates to end products, cell metabolic trajectories of *B. succiniciproducens* cells were predicted to move from the primary trajectory (multi-signal cells) to trajectory 1 (transposase$^+$ formate/nitrite transporter cells) and trajectory 2 (integrase$^+$ cells). Further pseudo-time differential gene analysis confirmed this transformation (Fig. 6c and Extended Data Fig. 9c).

To explore which genes contribute to which cell state transitions, we generated a pseudo-time heatmap for *B. succiniciproducens* cells and obtained 2,051 genes that co-varied across pseudo-time (Fig. 6c). Among these, 603 were enriched between the two trajectories and involved in 147 different pathways (Supplementary Table 9).

https://doi.org/10.1038/s41564-024-01723-9

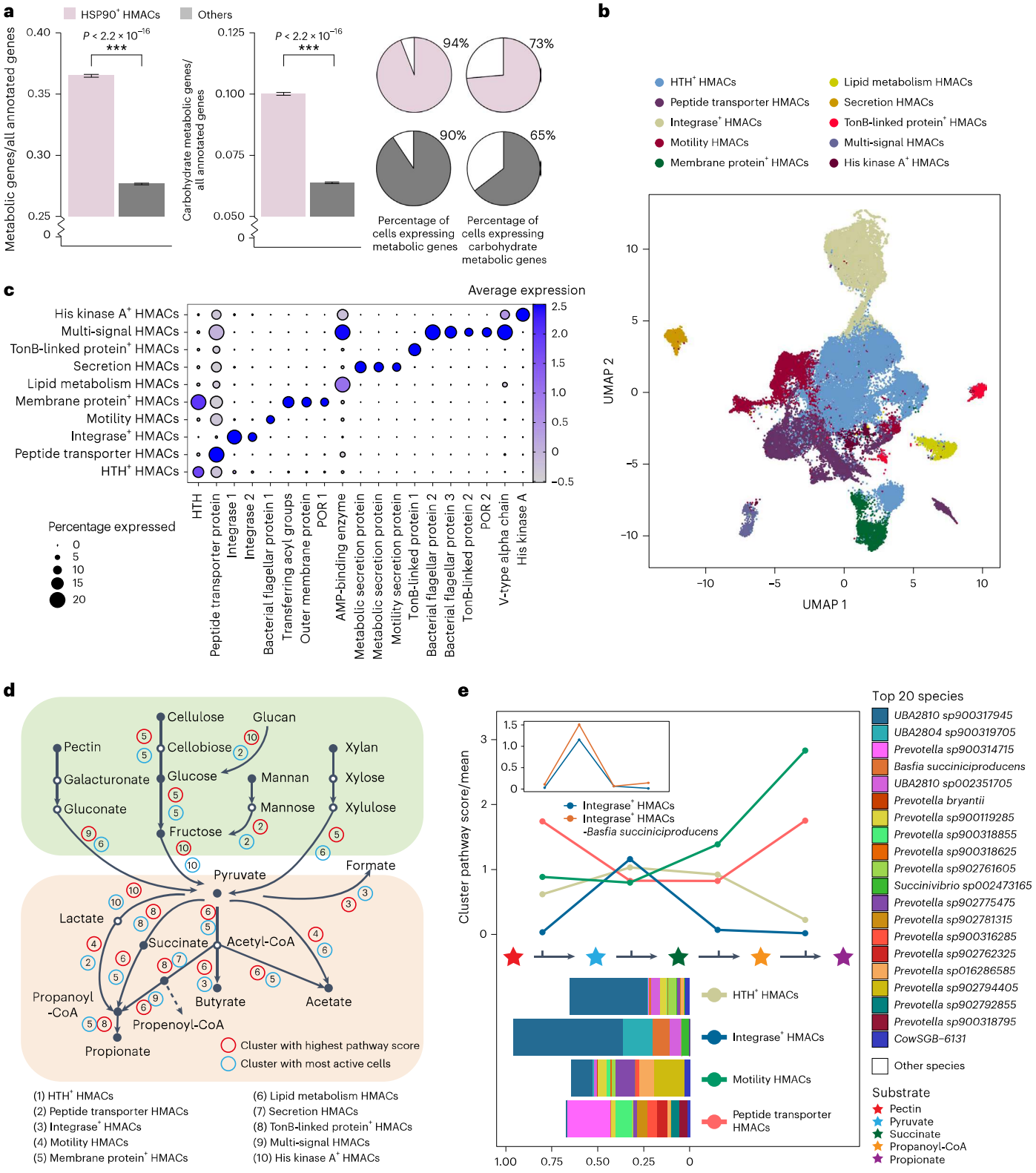


Fig. 5 | Metabolic insights into rumen classic carbohydrate pathways. **a**, Total metabolic FGPs (***$P < 2.2 \times 10^{-16}$) and carbohydrate metabolic FGPs (***$P < 2.2 \times 10^{-16}$) of HMACs and other cells (n = 49,854 and 123,014, respectively). The data are presented as mean values ± s.e.m. A two-sided Wilcoxon rank-sum test was used for data analysis. Adjustment was not made because there were no multiple comparisons. **b**, Sub-population functional clusters generated from re-clustered HMACs. **c**, Expression of the proteins encoded by the maker genes of HMAC sub-population functional clusters. **d**, Activities of HMAC sub-population functional clusters in rumen classic carbohydrate metabolism. The red circles represent the clusters with the highest FGPs, whereas the blue circles represent the clusters with the highest percentages of active cells. **e**, FGP change and species of four clusters in the four continuous steps in the metabolization of pectin to produce pyruvate. Top: The FGPs of the four clusters at each step. Bottom: The species composition of the four clusters. Inset: Comparison of the *B. succiniciproducens*' and integrase+ HMAC cells' average FPGs during the conversion of pyruvate to succinate. POR, Pyruvate:ferredoxin (flavodoxin) oxidoreductase.

https://doi.org/10.1038/s41564-024-01723-9

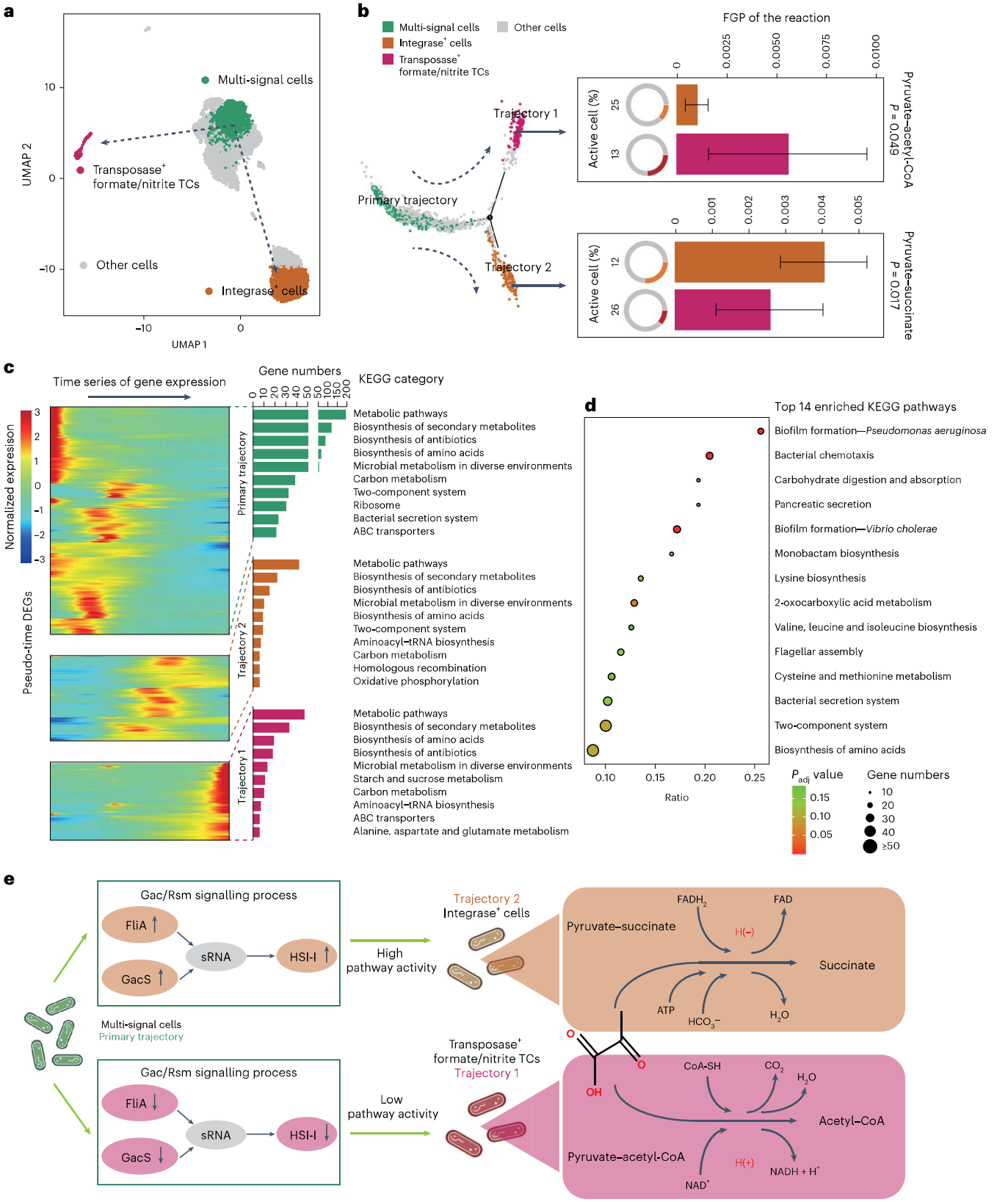


Fig. 6 | Functional heterogeneity and cellular trajectories of *B. succiniciproducens*. **a**, Sub-population functional clusters generated by re-clustering analysis of *B. succiniciproducens*. The black arrows indicate possible cell cluster transformation relationships identified in subsequent analyses. **b**, Pseudo-time analysis of three targeted subclusters and metabolic activities within two subclusters in two trajectories (For the Trajectory 1 and Trajectory 2, $n = 58$ and 108 biologically independent cells, respectively). The data are presented as mean values ± s.e.m. A two-sided Wilcoxon rank-sum test was used for data analysis. **c**, DEGs from pseudo-time analysis. The heatmap presents the time series of gene expression. The bar graph shows the numbers of DEGs per trajectory in the top ten KEGG pathways. **d**, KEGG functional enrichment analysis of all of the DEGs. A two-sided Fisher's exact test was used for data analysis. A Benjamini–Hochberg adjustment was made for multiple comparisons. **e**, Underlying mechanism of different cell fates within different metabolic activities. P_{adj}, adjusted P value; tRNA, transfer RNA. HSI-1, Hcp1 secretion island I; FliA, RNA polymerase sigma factor FliA.

The biofilm formation pathway (ko02025) showed the highest enrichment score (rich factor = 0.256; adjusted $P < 0.001$) owing to the significant changes of 21 genes (Fig. 6d). Specifically, the integrase$^+$ cells were significantly more active than the transposase$^+$ formate/nitrite transporter cells in the biofilm formation pathway ($P < 0.001$). Genes are mainly enriched in the Gac/Rsm signalling process of this pathway (Extended Data Fig. 9d and Supplementary Table 10), which is related to cytotoxicity, motility and cellular response to the environment[46]. The results revealed that the Gac/Rsm signalling process plays an essential role in functional cluster transformation of *B. succiniciproducens*. Interestingly, the two transformed functional clusters have obvious heterogeneity in the phenotype of metabolic processes using pyruvate as a substrate: transposase$^+$ formate/nitrite transporter cells are significantly more active in the pyruvate to acetyl-CoA metabolic processes ($P = 0.049$), whereas integrase$^+$ cells are significantly more active in the pyruvate to succinate metabolic processes ($P = 0.017$) (Fig. 6b,e and Supplementary Table 11). Pyruvate is a key intermediate in glucose metabolism, involved in the interconversion of sugars, fats and amino acids in the body[37]. Pyruvate metabolic processes are closely related to ruminal hydrogen metabolism, connecting the major nodes of the microbial fermentation process and influencing methanogenesis[47]. The metabolism of pyruvate to succinate reduces hydrogen production, which is of great value in rumen methane emission reduction studies. These results provide functional insights into the in situ microbial cell state transformation analysis of complex microbial ecosystems at the single-cell level. Under natural conditions, complex competition and environmental factors affect microbial cells, leading to changes in gene expression[48,49]. Therefore, cell state transformation based on the gene expression structure has great potential for regulating microbial cell metabolic phenotypes.

Discussion

In this study, we report on MscT technology, which enables high-throughput capture and annotation of microbial cells in complex communities with numerous unknown and unculturable species. Using MscT, we constructed a single-cell atlas of the rumen microbiome, covering 174,531 high-quality single-cell transcriptomes from 2,534 microbial species. Different microbial scRNA-seq approaches, such as PETRI-seq applied to *Escherichia coli MG1655* cells[14], microSPLiT applied to *Bacillus subtilis PY79* cells[15] and BacDrop applied to *Klebsiella pneumoniae MGH66* cells[16], have reported prokaryotic expression profiles. However, they have only been applied to fewer than five pre-known species in simple synthetic microbial communities and focused on localized biological processes rather than global ecosystems[50]. Our previous methods have been successfully applied in high-throughput microbial scRNA-seq[17]. The current version increases the efficiency for all species owing to updated random primers and its pre-indexes-based foundation. Microbial pangenome-based computational analysis is an effective solution for microbial transcriptomic annotation, and is theoretically well suited for any complex microbial environment. Integration of the above two important techniques in MscT shows great advantages and synergistic effects, enabling the detection and annotation of a larger number of expressed genes in each cell compared with other existing microbial scRNA-seq approaches.

Single-cell RNA-seq of the human gut microbiome is much easier because of well-established microbial reference genomes[5,6,51]. There is a severe lack of sufficiently sized genomic datasets in most other complex microbial environments, such as the guts of other animals, soil and oceans. Similarly, current studies on the rumen microbiome are limited by the lack of a comprehensive and high-quality collection of rumen microbial genomes, which serves as a prerequisite of this study. The BGMGM is a bovine gastrointestinal microbial genomic database with more than 47,000 genomes and 25 million genes. Notably, the number of de-replicated putative species-level genomes in BGMGM was remarkably close to the estimated number of microbial species in the rumen (13,572 versus 13,616)[32]. The larger number and higher quality of rumen microbial genomes[20,32] enable a more accurate understanding of the structure and functionality of microbial genomes in the rumen. The construction of BGMGM will not only contribute to the development and use of MscT but also expand the systematic and holistic understanding of microbial environments.

As the rumen hosts many microbes with high species richness and active microbial fermentation[18,52], the rumen microbiome serves as a good model for microbial ecosystem research. The first application of the landmark metagenomic binning technique was conducted on rumen microbes[4] and has attracted increasing attention in rumen microbial biology. Using MscT, we have identified functional clusters consisting of many different species of microbial cells, based on active gene expression structure and functional heterogeneity. The fingerprint produced by MscT records gene expression information for each cell, thus providing an approach to exploring functional heterogeneity in strains that are similar in evolutionary relationships. The single-cell transcriptomic landscape allowed us to accurately characterize the core functional clusters and key dynamic metabolic features of microbes during pyruvate-centred carbohydrate metabolism. Our research renews our knowledge of classic carbohydrate metabolism in the rumen and will lead to the construction of dynamic cellular metabolic maps of microbial ecosystems, which is a milestone in the exploration of active microbial functional states at a higher resolution.

In summary, our MscT approach demonstrates significant advancements in resolution, quality and accuracy, making it the appropriate scRNA-seq method for studying complex microbial communities. The functional clusters identified through MscT have expanded our understanding of microbial functional heterogeneity and its biological basis with unprecedented details. Despite our in-depth studies, we are still limited by the boundaries of current technology. For example, only few (25) Archaea cells were captured due to the cell wall composition, leading to insufficient insights into rumen methane emissions. Meanwhile, methods to separate functional clusters have not been developed, resulting in validation through synthetic communities still being a great challenge. Future studies should focus on promoting the widespread availability of these approaches to microbiome research by standardizing existing techniques, improving cell capture capability and developing functional cluster isolation methods.

Methods

Animals and rumen sample collection

The experimental protocol (protocol number: 12410) was approved by the Animal Use and Care Committee of Zhejiang University (Hangzhou, China) and the procedures were conducted based on the university's guidelines for animal research.

A total of 30 Holstein dairy cows with similar body weight and days in milk and under the same diet were selected from commercial dairy farms in the same area in Hangzhou (see Source Data Fig. 1 for the ingredients and nutrient composition of the total mixed ration fed to the cows). The rumen fluid of each cow was collected using oral stomach tubes, followed by centrifugation at 3,000*g* for 2 min at 4 °C. The supernatant was removed and the remaining biomass was collected. Then, the tubes were deposited in a liquid nitrogen tank, which was transported back to the laboratory and stored at −80 °C.

No statistical methods were used to pre-determine sample sizes, but we collected 174,531 high-quality cells in the rumen fluid, equating to a large sample size and cell number for microbial scRNA-seq. This sample size is considered sufficiently representative in rumen microbial ecosystem studies. Data collection and analysis were not performed blind to the conditions of the experiments.

Metagenomic sequencing and binning

DNA extraction, library construction and sequencing. The total DNA extraction of 30 rumen fluid samples was carried out using the E.Z.N.A.

Stool DNA Kit (Omega Bio-tek) following the protocol described by Yu and Morrison[53]. After completing genomic DNA extraction, 1% agarose gel electrophoresis was used for detection. For each sample, 1 µg genomic DNA was taken as the procedure template and sheared using a Covaris S220 Focused-ultrasonicator. The DNA was then fragmented into approximately 450 base pairs (bp) for library preparation. All of the samples were sequenced in an Illumina HiSeq X instrument with the 150-bp pair-end mode.

Pre-processing of raw sequencing data. To improve the quality, the raw data were trimmed using Trimmomatic version 0.36 (ref. 54) to remove adaptors and bases containing non-A, -G, -C and -T at the 5′ end, as well as any reads with a sequencing quality value of <20 or containing up to 10% N. After removing the adaptors and trimming, reads <75 bp in length were discarded. To further decrease the potential contamination from the host, the retained reads were mapped to the bovine genome from RefSeq (NCBI RefSeq assembly GCF_002263795.2) using the BWA mem algorithm (parameters: -M -k 32 -t 16; http://bio-bwa.sourceforge.net/bwa.shtml) and any aligned reads were removed. The remaining high-quality reads without host genome contamination were considered to be clean reads and were used for the further analysis.

Metagenomic binning. A set of contigs for each sample was generated using MegaHit version 1.1.1-2-g02102e1 with the parameters --min-contig-len 500 (ref. 55). MetaBAT2 version 2.11.1 (ref. 56) was used to perform binning on individual sample assemblies, and the completeness and contamination of all bins were obtained using CheckM version 1.1.3 (ref. 57). Bins with a completeness of ≥50% and a contamination of <10% were marked as filtered bins. Then, the bin abundance in each sample was quantified using the quant_bins module of metaWRAP version 1.3 (ref. 58).

The BGMGM

Microbial genomic datasets from public resources. We collected 54,403 bovine gastrointestinal microbial genomes (Supplementary Table 1) from 2,311 samples from ten different studies[8,19–27]. Overall, the samples covered 11 regions: China (Anhui, Hainan, Hangzhou, Henan, Hubei, Guangxi, Nanjing, Tibet and Yunnan), Scotland and Kenya. The metagenomes were sampled from ten gastrointestinal sites: the rumen, reticulum, omasum, abomasum, duodenum, jejunum, ileum, colon, caecum and rectum. The genomes from the rumen were the most dominant (over 29,255). All of the genomic data are available in public databases (PRJNA656389, PRJEB21624, PRJEB39057, PRJNA526070, PRJNA597489, PRJNA657455, PRJNA657473, PRJEB31266, PRJEB21624 and https://db.cngb.org/qtp).

De-redundancy and taxonomic analysis. In total, we collected 1,312 genomes generated from the present study and 54,403 genomes from public studies. These 55,715 bovine gastrointestinal microbial genomes were organized to construct the BGMGM. Considering the various genome filter criteria from different studies, genome completeness and contamination were re-estimated using CheckM version 1.1.3 (ref. 57). The quality of genomes was determined based on the MIMAG standard[59] and only those of high or medium quality were kept (for high quality, completeness > 90% and contamination < 5%; for medium quality, completeness ≥ 50% and contamination < 10%; Genomes from public studies were kept only if they meet the 'completeness − 5 × contamination ≥ 50' condition to ensure high data quality). The medium-quality genomes were separated into two groups based on their quality score (for relative-high quality parts, QS ≥ 75; for relative-low quality parts, QS < 75; QS = completeness − 5 × contamination + ln[N50 − contigs]) to obtain more detailed results when comparing with high-quality genomes[60,61]. The near-complete genomes were distinguished by a quality score > 100 from the high-quality group to select a highest-quality portion for visualization. Then, species-level clustering was performed using dRep version 3.2.0 (ref. 62) with the option -pa 0.95, which sets the average nucleotide identity to 95%. Subsequently, 13,572 species-level genome operational taxonomic units (gOTUs) were generated for downstream analysis. The taxonomy of genomes was classified using GTDB-Tk[63] with the database version R214 and the toolkit version 2.3.2. To infer the phylogenetic position of studied genomes, phylogenomic trees of near-complete genomes were reconstructed using PhyloPhlAn 3.0 (ref. 64). The phylogenomic tree was generated based on the marker genes in the PhyloPhlAn database (http://cmprod1.cibio.unitn.it/databases/PhyloPhlAn)[28] and visualized using iTOL[65]. All of the filtered genomes were visualized with genome size, completeness, contaminations, N50 contigs and scaffolds in a matrix and the 13,572 non-redundant genomes were uniformly named CowSGB-X.

Gene and functional annotation. Protein-coding genes were predicted using Prodigal version 2.6.3 (ref. 66) with the option -p meta. Non-redundant microbial gene catalogues were clustered using CD-HIT-EST (setting: -c 0.95 -G 0 -T 140 -n 5 -aS 0.9 -M 0) to eliminate redundant sequences. Protein sequences of non-redundant microbial gene catalogues were annotated using diamond version 2.0.4 (ref. 67) against the COG[68] database and HMMER version 3.3 (https://hmmer.org/) against the Kyoto Encyclopedia of Genes and Genomes (KEGG)[69] and CAZy[70] databases. The results of annotations in different databases were integrated. Genes identified in at least two databases were considered together to determine their functions.

Single-cell RNA-seq of the rumen microbiome

Cell suspension preparation. The thawed rumen fluid samples were resuspended in 15 ml 4% paraformaldehyde (PFA) and then dispersed through gentle vortexing. The dispersed solutions were filtered through a cascade of cell strainers with assorted pore sizes from 70 µm (43-10070-40; pluriSelect) to 10 µm (43-10010-40; pluriSelect) to remove impurities and acquire optimal single-cell suspension. The filtered solutions were centrifuged at 4,000*g* for 10 min at 4 °C. The supernatants were aspirated and microbial pellets were resuspended in 10 ml 4% PFA and then incubated overnight at 4 °C with 10 rpm rotational shaking.

Cell permeabilization. Following overnight fixation, PFA was discarded by centrifuging the cells at 4,000*g* for 10 min at 4 °C and the pellets were resuspended in 5 ml cold PBS-RI (that is, 1× phosphate-buffered saline (PBS) with 1 U µl^{-1} RNase inhibitor (N8080119; Invitrogen)). Cells were centrifuged again and resuspended in 1 ml pre-chilled 100 mM Tris-HCL-RI (pH 7) (that is, 100 mM Tris-HCL (pH 7) with 1 U µl^{-1} RNase inhibitor), followed by centrifugation at 4,000*g* for 5 min at 4 °C. Pellets were resuspended in 250 µl pre-chilled 0.04% Tween-20 (A600560; Sangon Biotech) in PBS and permeabilized on ice for 3 min. Centrifugation was then performed twice at 4,000*g* for 5 min at 4 °C and then the cells were counted. Approximately 50 million cells were obtained, resuspended in 200 µl lysozyme and then mixed on ice.

Cell wall digestion was set up with the reaction system consisting of 147.5 µl cells in nuclease-free water, 40 µl Lyso-Buffer, 2.5 µl RNase inhibitor and 10 µl lysozyme. The reagents were included in the VITApilote-PFT1200 kit (R20115124; M20 Genomics). Cells were incubated at 37 °C in a thermocycler for 15 min, then 1 ml pre-chilled PBS-RI was added to stop the incubation. Centrifugation was then performed twice at 4,000*g* for 5 min at 4 °C and the cells and pellets were resuspended in 1 ml pre-chilled PBS-RI. Cells were counted and a total of 5 million were prepared for ongoing processing.

In situ pre-barcoding reverse transcription. For the reverse transcription reaction, cells were evenly distributed into 14 PCR tubes and a pre-barcoded random primer was added to each. Each PCR tube was set up as follows: 2.25 µl cells in PBS, 0.25 µl 100 mM dNTPs, 1 µl 5× reverse transcription buffer, 0.25 µl RNase inhibitor, 0.25 µl reverse transcriptase and 1 µl 10 µM pre-barcoded random primer.

The reagents and primers were included in the VITApilote-PFT1200 kit. The PCR tubes were incubated with 12 cycles of multiple annealing ramping from 8–42 °C and then 42 °C for 30 min.

dA tailing. Following reverse transcription, 1 µl 50 mM EDTA was added to each PCR tube to terminate the incubation. Cells were transferred to one PCR tube and washed three times to deplete the residual reagent and random primers. The dA tailing reaction system was set up as follows: 39 µl cells in PBS, 5 µl buffer T2, 5 µl buffer T1, 0.5 µl TT enzyme and 0.5 µl 100 mM dATP. The reaction mix was incubated at 37 °C for 30 min. The reagents were included in the VITApilote-PFT1200 kit.

Single-cell droplet generation. Cells were counted and diluted with density gradient solution. Cells, 2× DNA extension reaction mix and ready-to-use hydrogel barcoded beads were encapsulated into droplets using the microfluidic platform VITAcruizer DP400 (E20000131; M20 Genomics) and chip (E20000131; M20 Genomics). All reagents for droplet generation were included in the VITApilote-PFT1200 kit. Droplets were then incubated at 37 °C for 1 h, 50 °C for 30 min, 60 °C for 30 min and 75 °C for 20 min.

Complementary DNA purification and amplification. Droplets were broken with perfluorooctane after the extension reaction. The aqueous phase was purified using AMPure XP beads (A63881; Beckman Coulter) and the purified complementary DNA (cDNA) was amplified by PCR reaction, followed by purification with AMPure XP beads and elution with nuclease-free water. The reagents and primers were included in the VITApilote-PFT1200 kit. The eluted cDNA was quantified by Qubit 4.0 fluorometer (Q33238; Invitrogen) and measured by 4200 TapeStation (G2991BA; Agilent).

Library preparation and sequencing. The VAHTS Universal DNA Library Prep Kit for Illumina V3 (ND607-03/04; Vazyme) was used for library construction. The cDNA was qualified by end-repair and adenylation reaction. The reaction mix containing 50 ng fragmented cDNA, end-repair buffer end-repair enzymes and nuclease-free water was incubated at 30 °C for 30 min and inactivated at 65 °C for 30 min. It was was then combined with ligation enzymes and a working adaptor and incubated at 20 °C for 15 min. The ligated DNA was purified with AMPure XP beads. Library amplification was performed, followed by purification. The final cDNA library was quantified by Qubit 4.0 and measured by 4200 TapeStation. Library sequencing was performed using the NovaSeq 6000 and S4 Reagent Kit with paired-end reads of 150 bp. The sequencing generated a total of ~3.8 Tb data from the 14 samples.

MscT analysis

Data quality control and filtering. Within the paired-end reads, forward reads (28 bp in total) containing the barcodes (20 bp) and unique molecular identifiers (8 bp) for distinguishing the single cells and genes were not trimmed, whereas the raw reverse reads were trimmed with Trimmomatic version 0.36 (ref. 54) with the options SLIDINGWINDOW:4:15 MINLEN:50. Clean reads shorter than 50 bp were discarded in further analyses. Each clean read was identified as belonging to a particular cell based on the barcodes. The number of valid cells in each sample was determined from the number of cellular reads. The thresholds (2,000, 2,500, 5,000 and 6,000) of cell reads were set depending on the quality, sequencing depth and number of cells required to extract valid cells (see the 10x standard; Supplementary Table 12).

Gene abundance calculation and high-quality single-cell screening. The clean reads from cells with taxonomic information were mapped to the non-redundant microbial gene catalogues from the BGMGM by BWA version 0.7.17-r1188 (ref. 71). The number of reads successfully matched was extracted from the alignment results using bedtools version 2.28.0. For each sample, the genes that covered fewer than three cells were ignored in further analysis. Then, all sample cells were integrated together. The cells with both n_{Count} and n_{Feature} values between three times the median absolute deviation were defined as high-quality cells and the remainder were removed. The selected high-quality cells with mapping results were combined and exported as the single-cell gene expression matrix.

Preparation of the Kraken2-based gOTUs database. For the taxonomic identification of each single cell, customized Kraken2-based gOTUs databases were constructed using Kraken2 (ref. 72). Ribosomal RNA genes were predicted using barrnap (https://github.com/tseemann/barrnap) and masked in gOTU genomes using bedtools (https://bedtools.readthedocs.io/). The kraken2-build module in the Kraken2 software package was used (settings: --no-masking --add-to-library) to create a classification database based on the masked gOTU genomes and corresponding taxonomic information. Then, the generated sequences were classified and constructed in the Kraken2 database. Furthermore, the count-kmer-abundances.pl script (setting: --read-length 100) was used to compute the *K*-mer pattern count of the database and store it as a db.kraken2.100mers.cnts file. Finally, the generate_kmer_distribution.py script (settings: -i db.kraken2.100mers.cnts -o KMER_DISTR.TXT) was used to generate a distribution file of the *K*-mer patterns.

Taxonomy determination. Clean reads for each single cell were classified by Kraken2 against the customized gOTUs databases. Cells were classified into seven phylogenetic levels (domain, phylum, class, order, family, genus and species) or unclassified. The normalized read number of taxonomies was calculated using Bracken (https://ccb.jhu.edu/software/bracken/), which used a Bayesian model to estimate normalized abundance. To obtain accurate taxonomic information for each cell, cells with more than 50% informative reads hitting the uppermost taxonomic level were considered to have a positive result (recognized as accurate annotated cells); otherwise, they were considered as the related taxonomic and were marked with the '_like' flag after the species name (Extended Data Fig. 3a). In this step, the unclassified reads were not calculated.

Filtering and benchmarking before clustering analysis. The single-cell gene expression matrix was imported into Seurat[73] (version 4.3.0) for subsequent analysis. Overall, cells with both $n_{\mathrm{Count\_RNA}}$ and $n_{\mathrm{Feature\_RNA}}$ values within threefold median absolute deviation were retained to be high quality and all others were removed. The DoubletFinder[74] package (version 2.0.3) was used to remove doublets. Then, the genes that covered fewer than three cells were removed. A series of benchmarks were used to determine the optimal parameters for dimension value and resolution value in the clustering analysis. The dimension value was determined by ElbowPlot, which shows the standard deviations of different principal components. When the points fall on a plateau (where the standard deviation does not change much) after an inflection point, the corresponding principal component is selected as the best possible dimension value. The resolution value was determined using the clustree[75] package (version 0.5.1), which generates clustering trees to interrogate clusters along with resolution increases. When cell clusters start to mix after a certain resolution, that resolution is selected as the best possible resolution value.

Cell clustering and functional cluster identification. After filtering and benchmarking, the dimensionality of all high-quality cells was reduced by uniform manifold approximation and projection. Batch effects between samples were removed using Harmony[76] (version 0.1.0). The clusters were identified using the FindClusters function (resolution = 0.3) of Seurat. Differentially expressed genes (DEGs) were determined using the FindAllMarkers function (average log (fold change) > 0.25, adjusted *P* value < 0.05 and per cent > 0.1) of Seurat.

Among the DEGs, those with annotated information and well-defined functional characteristics were identified as marker genes for each cell cluster. When there were multiple genes pointing to a certain function in a cell cluster (if present), the cluster was named after that function.

Interaction analysis. The cell units were categorized by the cell information of the functional cluster and species (named as cluster−species; for example, HSP90^{+} HMACs−*B. succiniciproducens*). The cell units found in more than three samples with a minimum of ten cells for each sample were extracted for interaction analysis. A total of 213 cell units were included in the sparse inverse covariance estimation for ecological association inference (SPIEC-EASI) analysis[77]. The SPIEC-EASI analysis was performed using the R package SpiecEasi version 1.0.7 with Meinshausen–Buhlmann's neighbourhood selection method. A total of 519 interactions were identified and visualized using Gephi version 0.10 with the Fruchterman Reingold layout.

Re-clustering and pathway functional activity analysis. The cells of the functional cluster HMACs and the species *B. succiniciproducens* were extracted for re-clustering analysis and the genes that covered fewer than three cells of these two groups were removed. The rest of the clustering and functional cluster identification analyses were the same as described previously. The functional activity of each cell for a certain pathway was calculated using the functional gene proportion (FGP, the number of functional genes in a certain pathway divided by the number of all annotated genes in a single cell). The genes used in the classic carbohydrate pathway were identified by the annotation results of the databases COG 2022-03, KEGG R107, Gene Ontology 2023-01-01 and CAZy 2022 (Supplementary Table 13).

Pseudo-time analysis. After re-clustering analysis, the cells of species *B. succiniciproducens* were extracted for pseudo-time analysis (Supplementary Table 14). The Monocle 2 (ref. 41) package (version 2.28.0) was used to discover cell functional state transformations. CellDataSet data were constructed from Seurat data using the function newCellDataSet. DEGs were calculated using the function differentialGeneTest. Genes with a q value of <0.01 were regarded as DEGs. The DEGs were sorted and imported into CellDataSet data using the function SetOrderingFilter. The pseudo-time trajectory was constructed using the DDRTree algorithm with default parameters. The dynamic expression changes of the determined DEGs were visualized using the plot_pseudotime_heatmap function.

Enrichment analysis. For functional enrichment, a two-tailed Fisher's exact test (see Source Data Fig. 6) was used to evaluate the enrichment of the DEGs against the non-redundant genes of the ruminant gastrointestinal microbiome. A corrected P value of <0.05 indicated significance (see the formula below)[78,79].

The false discovery rate (FDR) was calculated based on the nominal P value from the hypergeometric test and the following formula (P the P_{pathwayA}, value of KEGG pathway A; a, the number of DEGs in KEGG pathway A; b, the number of non-DEGs in KEGG pathway A; c, the number of DEGs in other KEGG pathways; d, the number of non-DEGs in other KEGG pathways; $n = a + b + c + d$; P_{FDR}, the false discovery rate of P_{pathwayA}; P_{length}, the total number of P values; P_{rank}, the rank number of P_{pathwayA}):

$$P\text{pathwayA} = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

$$P\text{FDR} = P\text{pathwayA} \times P\text{length}/P\text{rank}$$

Statistical analysis. The statistical significance of the differences in quality between high and medium quality and the differences in FGPs between cell clusters was analysed by Wilcoxon rank-sum test. The statistical significance of the inter-cluster differences in FGPs for COG pathways in the same species was analysed using the Kruskal–Wallis test. A P value of <0.05 was considered statistically significant. The inter-cluster multiple comparisons of FGPs for COG pathways in *B. succiniciproducens* were performed using Dunn post-hoc tests (for continuous variables, the R package FSA was used). The normality and equal variances were formally tested. The P values of the inter-cluster multiple comparisons and enrichment analysis were adjusted by the FDR (using the Benjamini–Hochberg method). An adjusted P value of <0.05 was considered statistically significant.

Reporting summary

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

Data availability

All of the raw sequencing data of the MscT have been deposited to the Genome Sequence Archive database with the accession number CRA012211. The genome files of MAGs in the BGMGM, gene annotation files and intermediate files resulting from quality control, benchmarking and other processes have been submitted to the Figshare database at https://figshare.com/articles/dataset/Microbiome_single-cell_transcriptomics_reveal_functional_heterogeneity_of_metabolic_niches_covering_more_than_2_500_species_in_the_rumen/24844344 (ref. 80). Source data are provided with this paper.

Code availability

The main codes and scripts from this study were uploaded to GitHub (https://github.com/J-MimgHui/MscT_codes).

References

1. Nguyen, C. L. et al. High-resolution analyses of associations between medications, microbiome, and mortality in cancer patients. *Cell* **186**, 2705–2718.e17 (2023).
2. Albertsen, M. et al. Long-read metagenomics paves the way toward a complete microbial tree of life. *Nat. Methods* **20**, 30–31 (2023).
3. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
4. Hess, M. et al. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467 (2011).
5. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
6. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
7. Valles-Colomer, M. et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* **614**, 125–135 (2023).
8. Stewart, R. D. et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
9. Royo-Llonch, M. et al. Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. *Nat. Microbiol.* **6**, 1561–1574 (2021).
10. Tian, L. et al. Deciphering functional redundancy in the human microbiome. *Nat. Commun.* **11**, 6217 (2020).
11. Windels, E. M. et al. Bacterial persistence promotes the evolution of antibiotic resistance by increasing survival and mutation rates. *ISME J.* **13**, 1239–1251 (2019).

12. Lloréns-Rico, V. et al. Single-cell approaches in human microbiome research. *Cell* **185**, 2725–2738 (2022).
13. Ojala, T. et al. Current concepts, advances, and challenges in deciphering the human microbiota with metatranscriptomics. *Trends Genet.* **39**, 686–702 (2023).
14. Blattman, S. B. et al. Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat. Microbiol.* **5**, 1192–1201 (2020).
15. Kuchina, A. et al. Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**, eaba5257 (2021).
16. Ma, P. et al. Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell* **186**, 877–891.e14 (2023).
17. Xu, Z. Droplet-based high-throughput single microbe RNA sequencing by smRandom-seq. *Nat. Commun.* **14**, 5130 (2023).
18. Mizrahi, I., Wallace, R. J. & Moraïs, S. The rumen microbiome: balancing food security and environmental impacts. *Nat. Rev. Microbiol.* **19**, 553–566 (2021).
19. Seshadri, R. et al. Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36**, 359–367 (2018).
20. Xie, F. et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* **9**, 137 (2021).
21. Tong, F. et al. The microbiome of the buffalo digestive tract. *Nat. Commun.* **13**, 823 (2022).
22. Stewart, R. D. et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).
23. Wilkinson, T. et al. 1200 high-quality metagenome-assembled genomes from the rumen of African cattle and their relevance in the context of sub-optimal feeding. *Genome Biol.* **21**, 229 (2020).
24. Xue, M.-Y. et al. Investigation of fiber utilization in the rumen of dairy cows based on metagenome-assembled genomes and single-cell RNA sequencing. *Microbiome* **10**, 11 (2022).
25. Li, X. et al. A unified catalog of 19,251 non-human reference species genomes provides new insights into the mammalian gut microbiomes. Preprint at *BioRxiv* https://doi.org/10.1101/2022.05.16.491731 (2022).
26. Solden, L. M. et al. Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).
27. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
28. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
29. Watson, M. New insights from 33,813 publicly available metagenome-assembled-genomes (MAGs) assembled from the rumen microbiome. Preprint at *BioRxiv* https://doi.org/10.1101/2021.04.02.438222 (2021).
30. Louca, S. et al. Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* **2**, 936–943 (2018).
31. Escalas, A. et al. Microbial functional diversity: from concepts to applications. *Ecol. Evol.* **9**, 12000–12016 (2019).
32. Tikhonov, M. Theoretical microbial ecology without species. *Phys. Rev. E* **96**, 032410 (2017).
33. Taxis, T. M. et al. The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. *Nucleic Acids Res.* **43**, 9600–9612 (2015).
34. Wang, M. et al. Even allocation of benefits stabilizes microbial community engaged in metabolic division of labor. *Cell Rep.* **40**, 111410 (2022).
35. Wu, G. et al. Two competing guilds as a core microbiome signature for health recovery. Preprint at *BioRxiv* https://doi.org/10.1101/2022.05.02.490290 (2022).
36. Liu, H. et al. Ecological dynamics of the gut microbiome in response to dietary fiber. *ISME J.* **16**, 2040–2055 (2022).
37. Stacpoole, P. W. & McCall, C. E. The pyruvate dehydrogenase complex: life's essential, vulnerable and druggable energy homeostat. *Mitochondrion* **70**, 59–102 (2023).
38. Sun, H.-Z. et al. Multi-omics reveals functional genomic and metabolic mechanisms of milk production and quality in dairy cows. *Bioinformatics* **36**, 2530–2537 (2020).
39. Cimini, D. et al. Improved production of succinic acid from *Basfia succiniciproducens* growing on *A. donax* and process evaluation through material flow analysis. *Biotechnol. Biofuels* **12**, 22 (2019).
40. Kuhnert, P. et al. *Basfia succiniciproducens* gen. nov., sp. nov., a new member of the family Pasteurellaceae isolated from bovine rumen. *Int. J. Syst. Evol. Microbiol.* **60**, 44–50 (2010).
41. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
42. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
43. Wang, Q. et al. Prediction of prokaryotic transposases from protein features with machine learning approaches. *Microb. Genom.* **7**, 000611 (2021).
44. Atkovska, K. et al. Energetics and mechanism of anion permeation across formate-nitrite transporters. *Sci. Rep.* **7**, 12027 (2017).
45. Maertens, G. N. et al. Structure and function of retroviral integrase. *Nat. Rev. Microbiol.* **20**, 20–34 (2022).
46. Latour, X. The evanescent GacS signal. *Microorganisms* **8**, 1746 (2020).
47. Li, Q. S. et al. Dietary selection of metabolically distinct microorganisms drives hydrogen metabolism in ruminants. *ISME J.* **16**, 2535–2546 (2022).
48. Foster, K. R. et al. Competition, not cooperation, dominates interactions among culturable microbial species. *Curr. Biol.* **22**, 1845–1850 (2012).
49. Friedman, N. et al. Compositional and functional dynamics of the bovine rumen methanogenic community across different developmental stages. *Environ. Microbiol.* **19**, 3365–3373 (2017).
50. Moraïs, S. et al. The road not taken: the rumen microbiome, functional groups, and community states. *Trends Microbiol.* **27**, 538–549 (2019).
51. Nayfach, S. et al. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
52. Mizrahi, I. et al. Review: the compositional variation of the rumen microbiome and its effect on host performance and methane emission. *Animal* **12**, s220–s232 (2018).
53. Yu, Z. & Morrison, M. Improved extraction of PCR-quality community DNA from digesta and fecal samples. *BioTechniques* **36**, 808–812 (2004).
54. Bolger, A. et al. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
55. Li, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).
56. Kang, D. D. et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
57. Parks, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

58. Uritskiy, G. V. et al. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
59. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
60. Orakov, A. et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol.* **22**, 178 (2021).
61. Zeng, S. et al. A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat. Commun.* **13**, 5139 (2022).
62. Olm, M. R. et al. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
63. Chaumeil, P.-A. et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
64. Asnicar, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
65. Letunic, I. et al. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
66. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
67. Buchfink, B. et al. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
68. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
69. Kanehisa, M. et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
70. Drula, E. et al. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).
71. Li, H. et al. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
72. Lu, J. et al. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).
73. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
74. McGinnis, C. S. et al. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337.e4 (2019).
75. Zappia, L. et al. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience* **7**, giy083 (2018).
76. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
77. Kurtz, Z. D. et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
78. Bu, D. et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325 (2021).
79. Klopfenstein, D. V. et al. GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
80. Minghui, J. Microbiome single-cell transcriptomics reveal functional heterogeneity of metabolic niches covering more than 2,500 species in the rumen. *Figshare* https://doi.org/10.6084/m9.figshare.24844344.v1 (2024).

Acknowledgements

We thank all of the members of the Institute of Dairy Science, College of Animal Sciences, Zhejiang University for assistance with sample collection. This work was supported by the National Natural Science Foundation of China (grant no. 32322077 to H.-Z.S.), National Key R&D Program of China (grant no. 2022YFD1301700 and 2023YFE0123100 to H.-Z.S.), Natural Science Foundation of Zhejiang Province (grant no. LR23C170001 to H.-Z.S.) and Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (grant no. 2021R01012 to Y.W.).

Author contributions

H.-Z.S. supervised the project and designed the research. H.-Z.S., M.J., S.Z., Y. Tang, X.L. and Y. Tao constructed the BGMGM. M.J., S.Z., Y. Tang and X.L. collected the rumen fluid samples. M.S. performed the RNA-seq experiments with assistance from Y.W. M.J., S.Z., T.Z. and Y. Tao performed the species annotation and gene alignment at the single-cell level. M.J. and S.Z. performed the cluster analysis, cell-type annotation, pathway analysis and pseudo-time analysis. M.J., S.Z., H.C. and H.-Z.S. interpreted the data. M.J., S.Z. and J.X. visualized the results. M.J., S.Z. and M.-Y.X. wrote the paper. H.-Z.S., Y.W. and J.-X.L. revised the paper. All authors read and approved the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

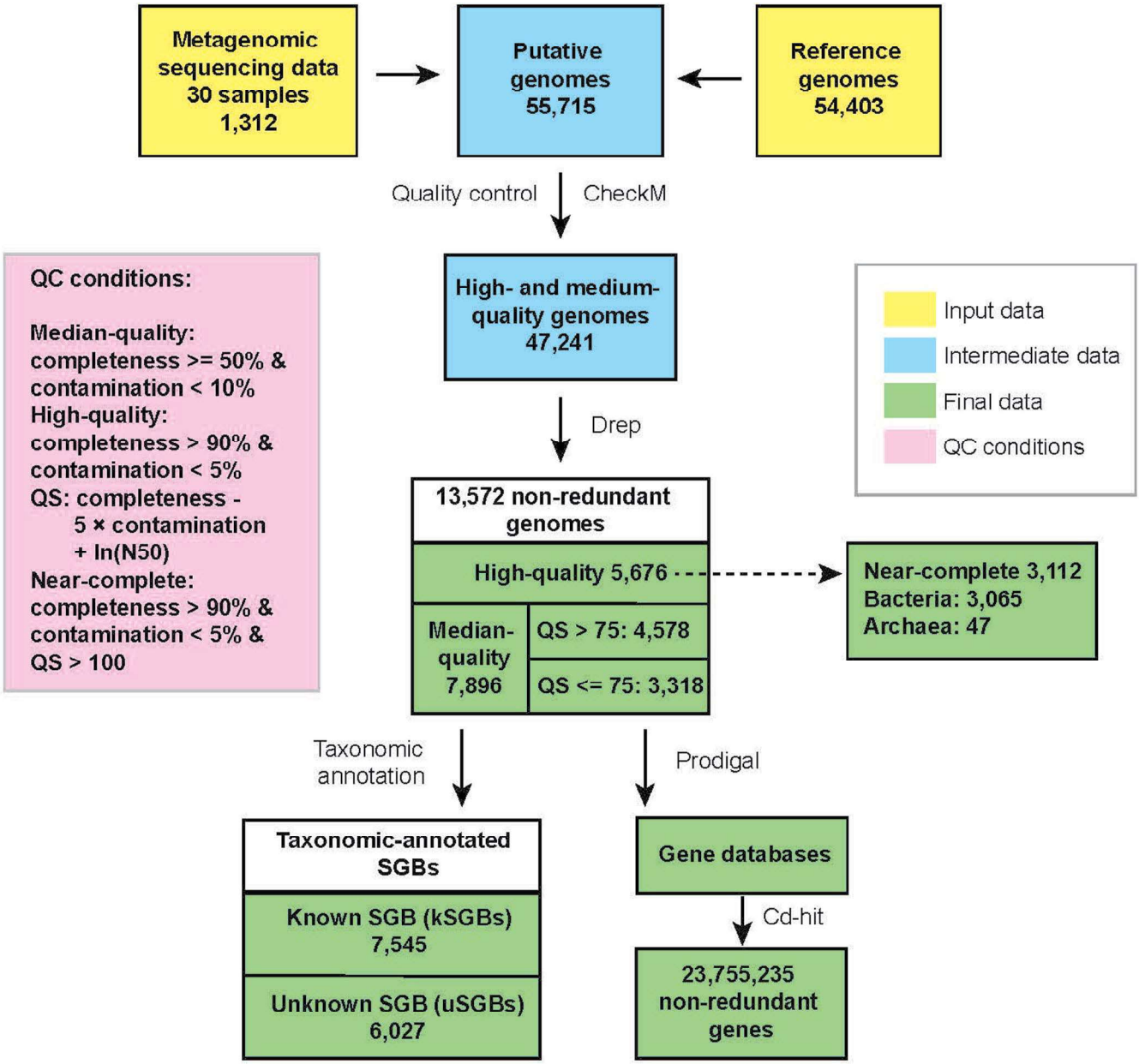
Extended data is available for this paper at https://doi.org/10.1038/s41564-024-01723-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-024-01723-9.

Correspondence and requests for materials should be addressed to Yongcheng Wang or Hui-Zeng Sun.

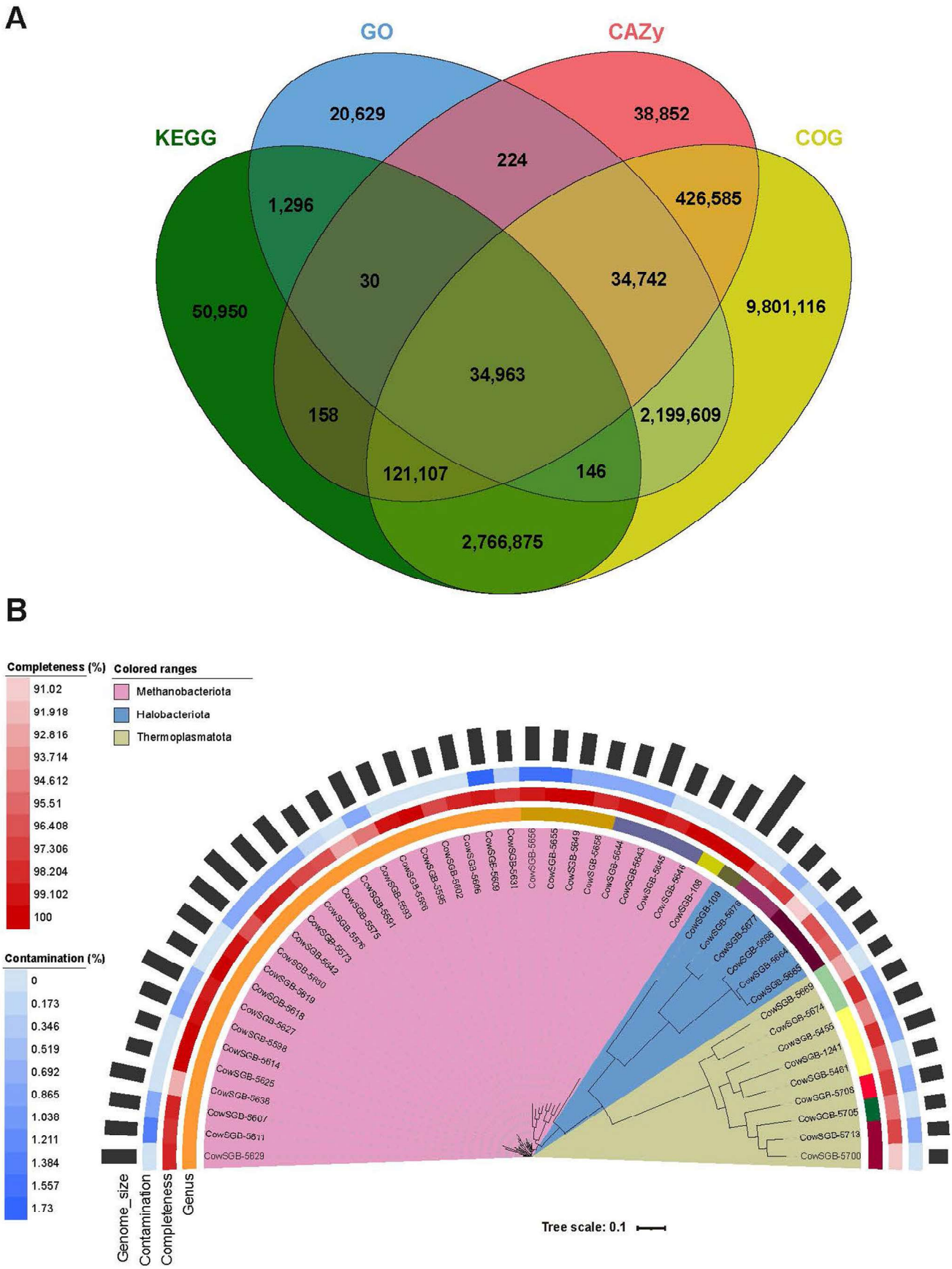
Peer review information *Nature Microbiology* thanks Karthik Raman, James Volmer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-024-01723-9



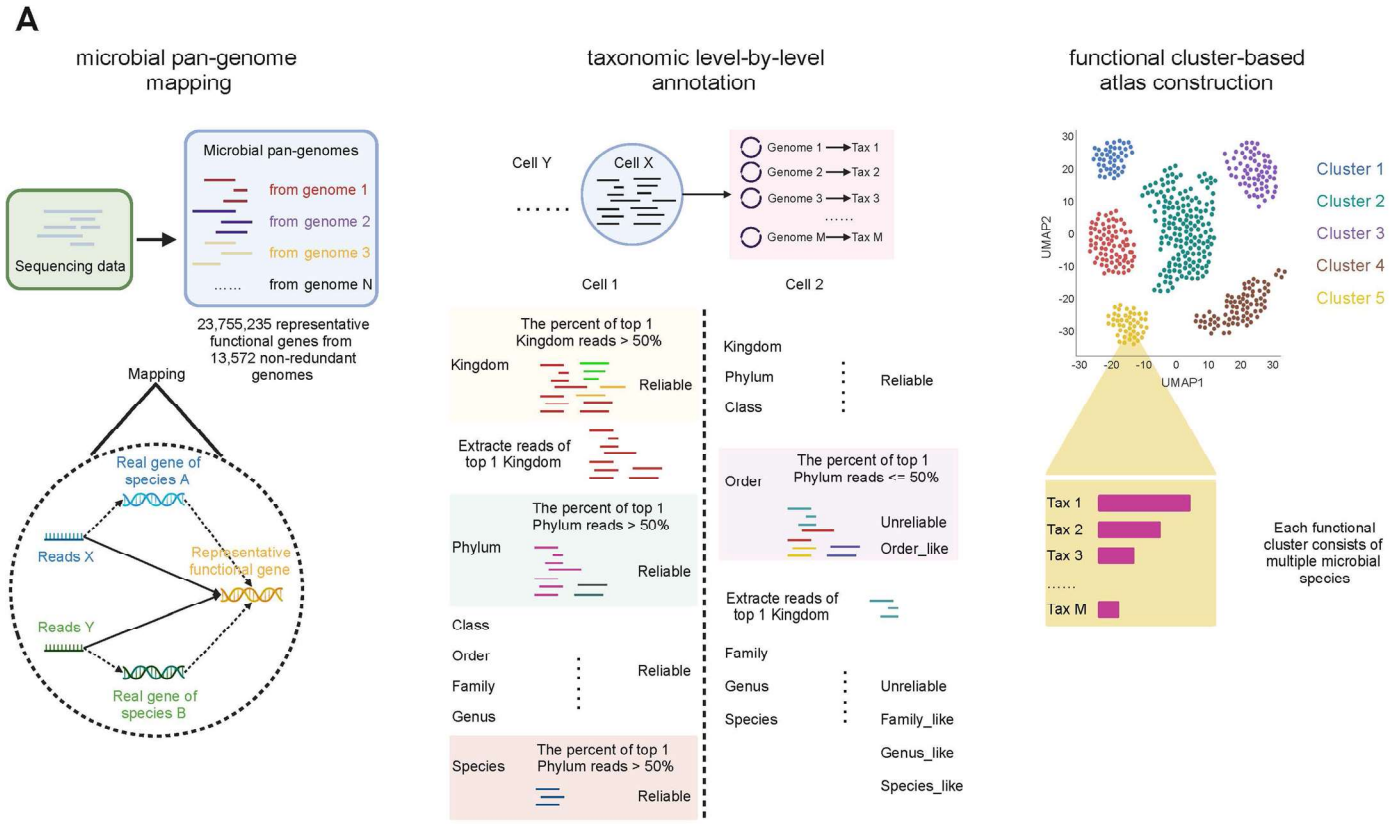
Extended Data Fig. 1 | Overall workflow of BGMGM construction. Workflow for the construction of the Bovine Gastro Microbial Genome Map (BGMGM).

https://doi.org/10.1038/s41564-024-01723-9



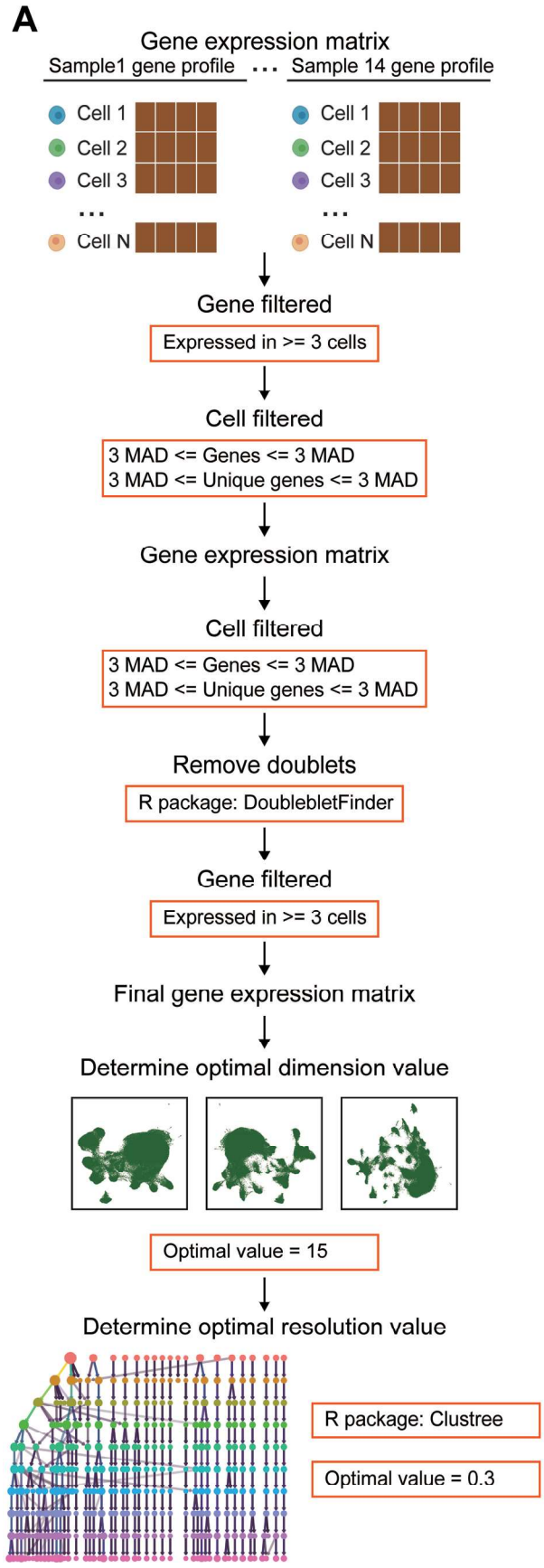
Extended Data Fig. 2 | The venn plot of annotated genes and the phylogenic tree of 47 Archaea MAGs. (**A**) The Venn plot of BGMGM genes annotated by KEGG database, GO database, CAZy database, and COG database. (**B**) The phylogenic tree of 47 Archaea MAGs. MAGs: metagenome assembled genomes. BGMGM, bovine gastro microbial genome map.

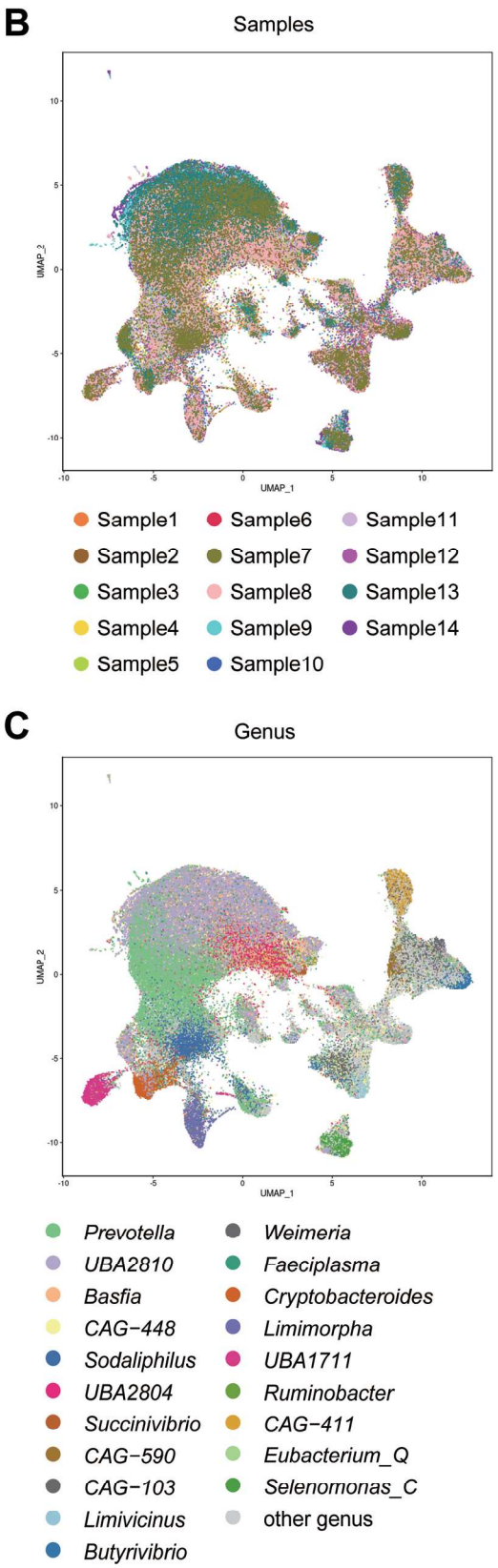
https://doi.org/10.1038/s41564-024-01723-9



Extended Data Fig. 3 | Microbiome single-cell transcriptomics computational analysis pipeline and performance. (**A**) Computational analysis pipeline including microbial pan-genome mapping, taxonomic level-by-level annotation, and functional cluster-based atlas construction. (**B**) The UMI numbers and unique gene numbers in each sample. UMI, unique molecular identifiers. Each box represents the interquartile range (IQR), in which the middle line represents the median. The whiskers extend to 1.5 × IQR.

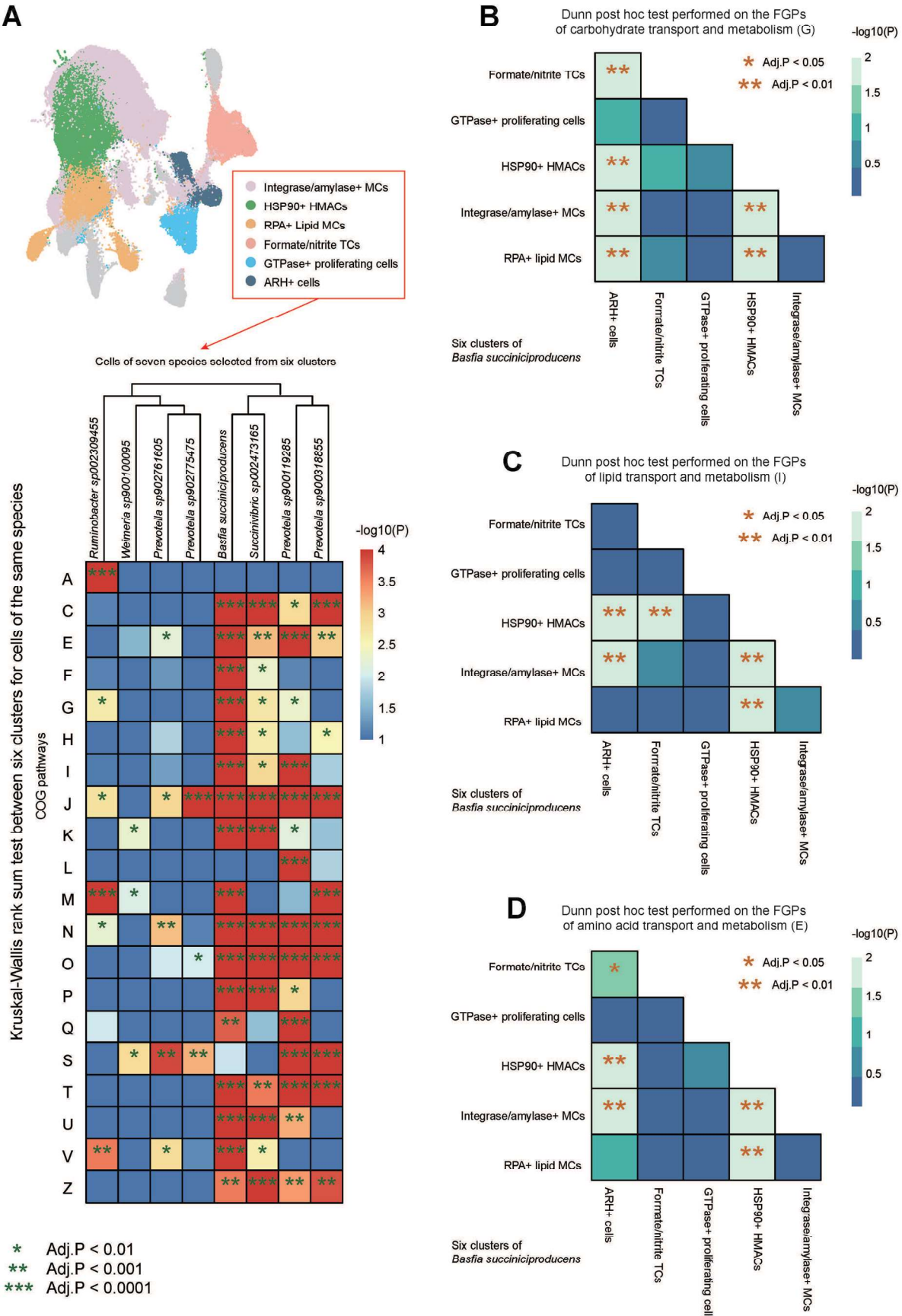
https://doi.org/10.1038/s41564-024-01723-9





Extended Data Fig. 4 | Heterogeneity among different functional clusters and species. (**A**) The cell and gene filtering steps as well as the benchmarking processes to determine the dimension and resolution values. (**B**) The UMAP plots for cells of different samples. (**C**) The UMAP plots for cells of different genera. UMAP, Uniform Manifold Approximation and Projection.

https://doi.org/10.1038/s41564-024-01723-9

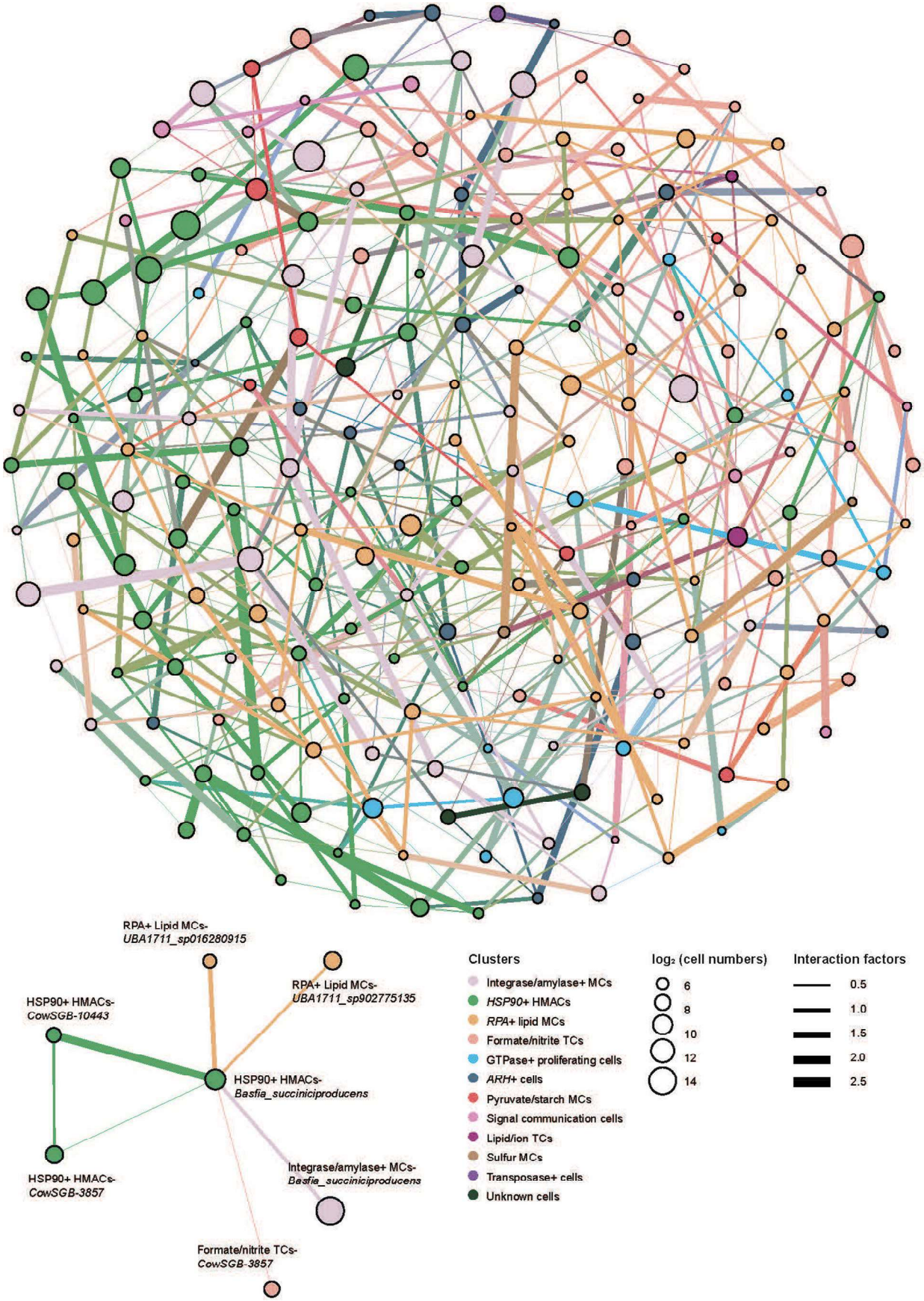


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41564-024-01723-9

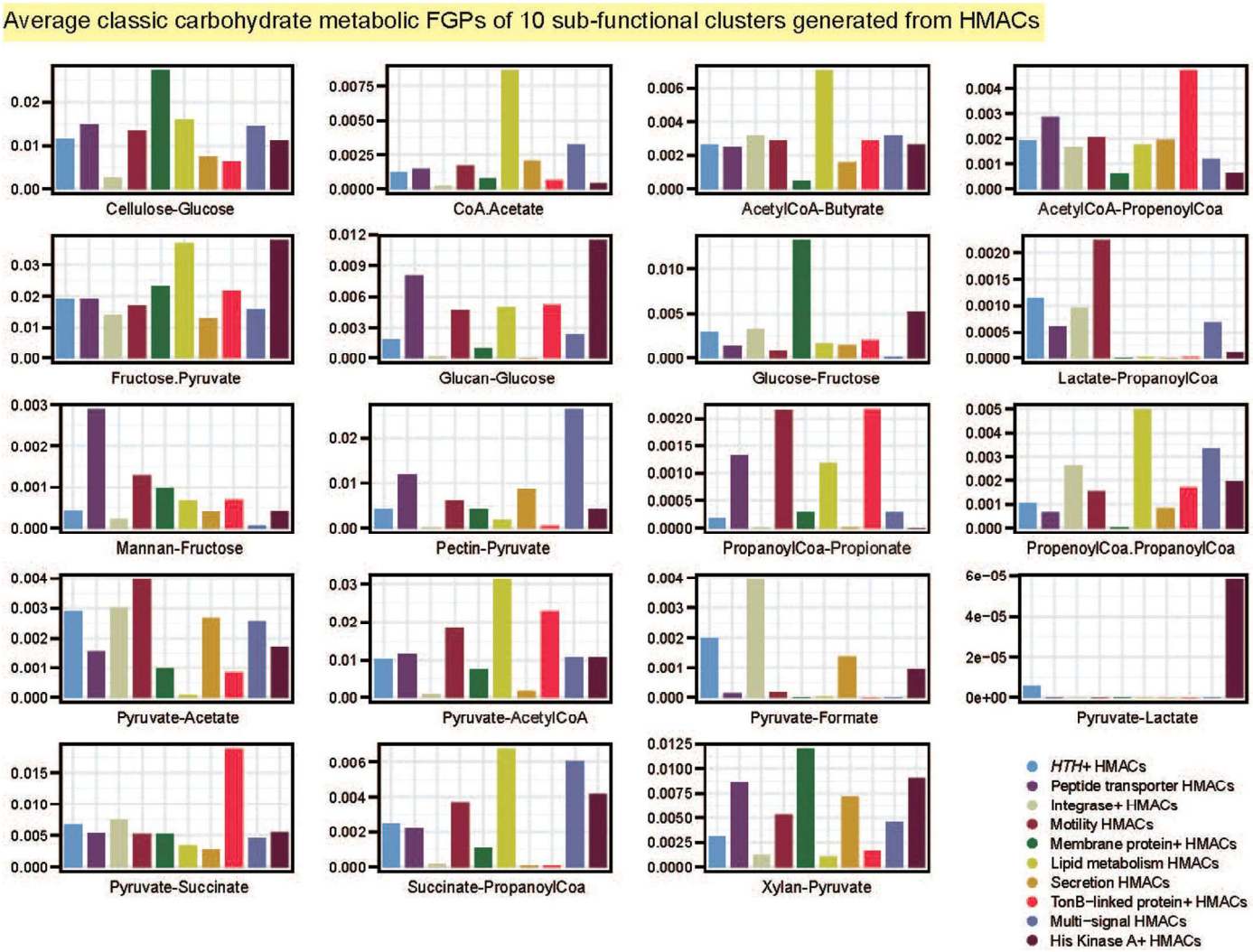
Extended Data Fig. 5 | Differences between six functional clusters in the same species analyzed by the Kruskal–Wallis test with Dunn post hoc tests. (**A**) The P values of inter-cluster comparison for FGPs in eight species. The UMAP plot presented the functional clusters involved in the analysis. The heat map showed the P values. (**B**) Dunn post hoc test performed on the FGPs of carbohydrate transport and metabolism. (**C**) Dunn post hoc test performed on the FGPs of lipid transport and metabolism. (**D**) Dunn post hoc test performed on the FGPs of amino acid transport and metabolism. FGP: functional gene proportion (the number of functional genes in a certain pathway/the number of all annotated genes in single cell); UMAP, Uniform Manifold Approximation and Projection.

https://doi.org/10.1038/s41564-024-01723-9



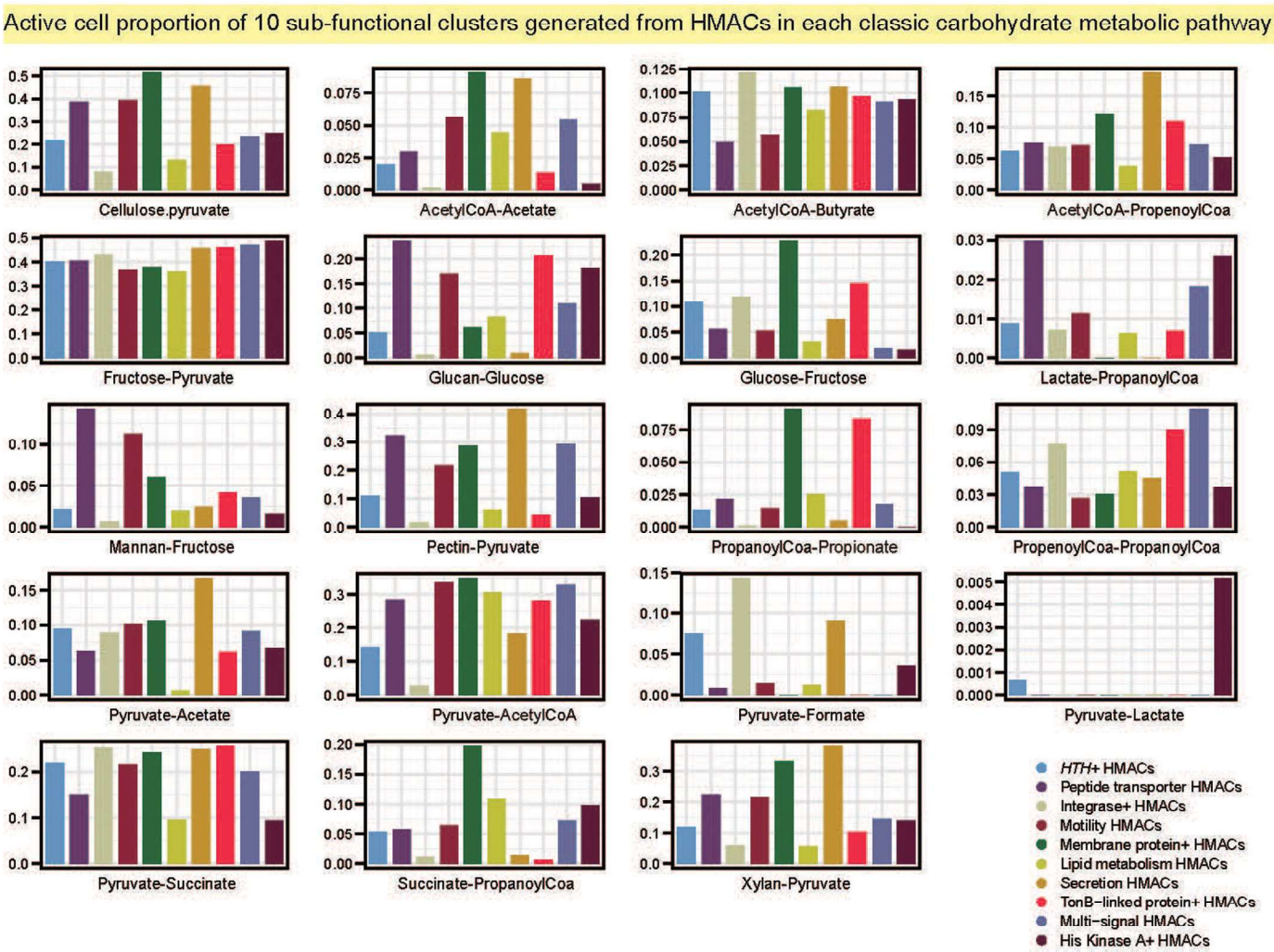
Extended Data Fig. 6 | SPIEC-EASI analysis. The interaction networks of 213 cell units and the interactions between the HSP90^{+} HMACs—*Basfia_succiniciproducens* and other associated cell units. SPIEC-EASI, Sparse Inverse Covariance Estimation for Ecological Association Inference.

https://doi.org/10.1038/s41564-024-01723-9



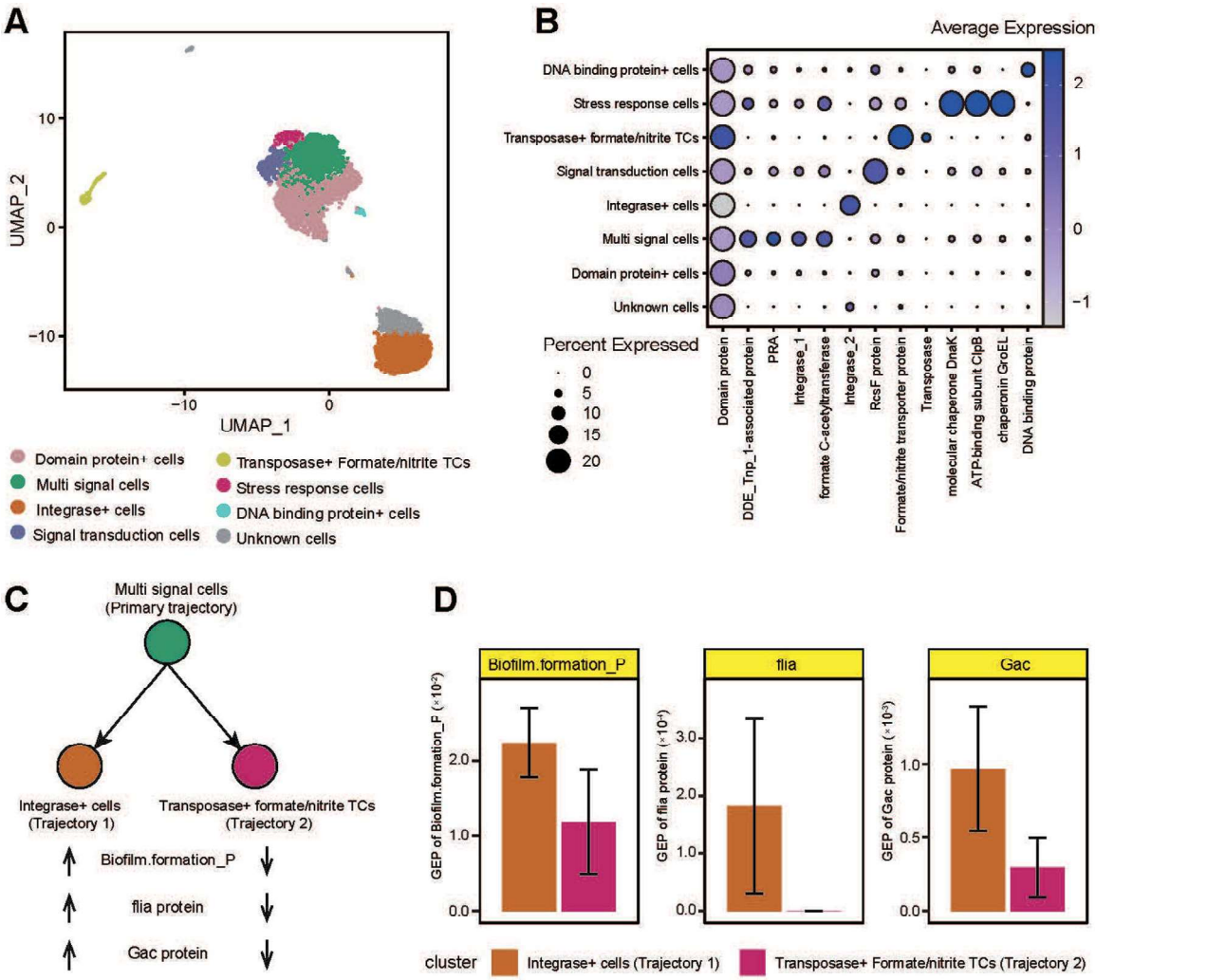
Extended Data Fig. 7 | Carbohydrate metabolic activity analysis. Average classic carbohydrate metabolic FGPs of 10 sub-functional clusters generated from HMACs. FGPs, functional gene proportions, the number of functional genes in a certain pathway/the number of all annotated genes; HMACs, high metabolic activity cells.

https://doi.org/10.1038/s41564-024-01723-9



Extended Data Fig. 8 | Active cell proportion analysis. Active cell proportion of 10 sub-functional clusters generated from HMACs in each classic carbohydrate metabolic pathway. HMACs, high metabolic activity cells.

https://doi.org/10.1038/s41564-024-01723-9



Extended Data Fig. 9 | Marker genes and biofilm formation pathway activity analysis of 8 sub-population functional clusters form *B. succiniciproducen* cells. (**A**) The UMAP plot of eiight sub-population functional clusters form *B. succiniciproducen* cells. (**B**) Marker genes of eight sub-population functional clusters form *B. succiniciproducen* cells. (**C**) Transformational relationships between clusters "Multi signal cells", "Integrase+ cells", and "Transposase+ formate/nitrite TCs". (**D**) "Biofilm.formation_P" pathway activity and two key gene proportion, n = 200 and 121 biologically independent cells. Data are presented as mean values +/- SEM. Two-side Wilcoxon rank sum test was used for data analysis. UMAP, Uniform Manifold Approximation and Projection.

nature portfolio

Corresponding author(s): Hui-Zeng Sun; Yongcheng Wang

Last updated by author(s): Feb 6, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All the public genomic data collected in this study are available in public databases (PRJNA656389, PRJEB21624, PRJEB39057, PRJNA526070, PRJNA597489, PRJNA657455, PRJNA657473, PRJEB31266, PRJEB21624, https://db.cngb.org/qtp).
Data analysis	Metagenomic sequencing and binning: The raw data was trimmed using Trimmomatic v.0.36 to remove adaptors and bases; The retained reads were mapped to the bovine genome from RefSeq (NCBI RefSeq assembly GCF_002263795.2) by BWA mem algorithm (parameters: -M -k 32 -t 16, http://bio-bwa.sourceforge.net/bwa.shtml) and the matched ones were removed; Clean reads were generated a set of contigs of each sample using MegaHit v.1.1.1-2-g02102e1; MetaBAT2 v.2.11.1 was applied to do binning; The completeness and contamination of all bins were obtained by CheckM v.1.1.3, with the completeness >= 50% and contamination < 10% ones being marked as “filtered bins”; The bin abundance in each sample was quanted by “quant_bins” module of metaWRAP v.1.3; Bovine Gastro Microbial Genome Map: Completeness and contamination of the public genomes were re-estimated using CheckM v1.1.3; For comparison, genome quality was estimated using CheckM2 v1.0.1; The taxonomy of genomes was classified using GTDB-Tk with the database version of R214_v2.3.2; Phylogenomic trees of near complete genomes were reconstructed by PhyloPhlAn 3.0; The phylogenomic tree was rooted according to the GTDB database and visualized using iTOL; Protein-coding genes were predicted with the Prodigal v2.6.3; Nonredundant microbial gene catalogs were clustered by cd-hit-est; Protein sequences of nonredundant microbial gene catalogs were annotated using diamond v2.0.4;

Single-cell RNA sequencing of rumen microbiome:
The raw reverse reads were trimmed with Trimmomatic v0.36;
The clean reads from cells with taxonomic information were mapped to the nonredundant microbial gene catalogs from BGMGM by BWA v0.7.17-r1188;
The number of reads successfully matched were extracted from the alignment results using bedtools v2.28.0;
For the taxonomic identification for each single cell, customized kraken2-based gOTUs databases were constructed by kraken2;
rRNA genes were predicted by barrnap (https://github.com/tseemann/barrnap) and masked in gOTU genomes using bedtools (https://bedtools.readthedocs.io/);
The normalized read number of taxonomies were calculated by Bracken (https://ccb.jhu.edu/software/bracken/);
The single-cell gene expression matrix was imported into Seurat (version 4.3.0);
The DoubletFinder package (version 2.0.3) was used to remove doublets;
Batch effects between samples were removed by Harmony (version 0.1.0);
The clusters were identified by the "FindClusters" function (resolution = 0.2) of Seurat;
The marker genes were determined by the "FindAllMarkers" function (|'avg_logFC'| > 0.25, 'p_val_adj' < 0.05 and pct > 0.1) of Seurat;
Monocle 2 package (version 2.28.0) was used to discover the cell functional state transformations;
The cds data was constructed from Seurat data by function "newCellDataSet";
The differential expression genes (DEGs) were calculated by function "differentialGeneTest";
The DEGs were Sorted and imported into cds data by function "SetOderingFilter";
The pseudotime trajectory was constructed by "DDRTree" algorithm with default parameters;
The dynamical expression changes of the determined DEGs were visualized by "plot_pseudotime_heatmap" function

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the raw sequencing data of the MscT has been deposited to Genome Sequence Archive (GSA) database with accession number CRA012211. The genome files of MAGs in BGMGM, gene annotation files, and intermediate files resulting from QC, benchmarking, and other processes have been submitted to the figshare database (https://figshare.com/s/5148ea8f39ca6733e051).The above data will be public along with the manuscript published. The main codes and scripts in this study were uploaded to Github (https://github.com/J-MimgHui/MscT_codes).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The study did not involve human participants, their data, or biological material.
Reporting on race, ethnicity, or other socially relevant groupings	The study did not involve human participants, their data, or biological material.
Population characteristics	The study did not involve human participants, their data, or biological material.
Recruitment	The study did not involve human participants, their data, or biological material.
Ethics oversight	The study did not involve human participants, their data, or biological material.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 30 Holstein dairy cows with similar body weight and days in milk were selected from commercial dairy farms. Fourteen rumen fluid samples collected from these cows were used for microbial single-cell transcriptome sequencing. We collected 174,531 high-quality cells in

	the rumen fluid, which is a large sample size and cell number for microbial single-cell RNA sequencing. Sufficiently representative in rumen microbial ecosystem studies.
Data exclusions	No data were excluded from the analyses.
Replication	All attempts at replication were successful.
Randomization	For the 30 cows in this study, we did not treat them additionally before and during the experiment. There was no need to divide the sample into different groups for this study, so the randomization is not relevant to this study.
Blinding	In this experiment, rumen fluid was collected from 30 Holstein dairy cows, which were not treated. Meanwhile, the animals included in the study were not grouped among this study. Therefore, the blinding was not relevant to our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	The study did not involve laboratory animals.
Wild animals	The study did not involve wild animals.
Reporting on sex	The 30 Holstein dairy cows in this study are all female. In the dairy farming industry, only female cows produce milk, so we select only females to ensure that our results are biologically significant enough to guide the industry.
Field-collected samples	A total of 30 Holstein dairy cows with similar body weight and days in milk were selected from commercial dairy farms in the same area in Hangzhou under the same diet. Rumen fluid of each cow was collected using oral stomach tubes, followed with centrifugation at 3,000 × g for 2 min at 4 °C. The supernatant was removed, and the remaining biomass was collected. Then the tubs were deposited in the liquid nitrogen tank transporting back to the laboratory where it got stored at -80 °C.
Ethics oversight	The experimental protocol (protocol number: 12410) was approved by the Animal Use and Care Committee of Zhejiang University (Hangzhou, China) and the procedures were conducted based on the university's guidelines for animal research.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*