2025, 65(1): 402-415
CSTR: 32112.14.j.AMS.20240483
DOI: 10.13343/j.cnki.wsxb.20240483

微生物学报
Acta Microbiologica Sinica
http://journals.im.ac.cn/actamicrocn

Data Paper 数据论文

基于 Kraken2 扩展标准数据库对反刍动物消化道微生物分类能力

翁玉楠[#]，甄永康[#]，王梦芝，王洪荣[*]

扬州大学 动物科学与技术学院，江苏 扬州 225009

翁玉楠, 甄永康, 王梦芝, 王洪荣. 基于 Kraken2 扩展标准数据库对反刍动物消化道微生物分类能力[J]. 微生物学报, 2025, 65(1): 402-415.
WENG Yunan, ZHEN Yongkang, WANG Mengzhi, WANG Hongrong. Classification ability of extended Kraken2 standard database for digestive tract microbiota in ruminants[J]. Acta Microbiologica Sinica, 2025, 65(1): 402-415.

摘　要：宏基因组学技术的应用丰富了对动物消化道中微生物组成以及功能的认识。当前，基于宏基因组测序读长(reads)水平的物种组成的分类比对水平普遍在 15%–45%。因此，提高宏基因组测序 reads 水平微生物的比对率，可进一步挖掘宏基因数据中的微生物信息。【目的】通过扩展 Kraken2 标准数据库来提高反刍动物消化道微生物的分类能力，从而进一步挖掘宏基因组数据中的微生物信息。【方法】本研究共收集了来自牛、绵羊和山羊瘤胃液、粪便以及消化道中 14 827 个宏基因组组装基因组(metagenome-assembled genomes, MAGs)，经质控过滤后，保留了 3 095 个物种级基因组箱(species-level genome bins, SGBs)，经物种分类以及功能预测后，SGBs 被整合进 Kraken2 标准数据库，并对其分类效果予以评估。【结果】在 SGBs 在基因组分类数据库(genome taxonomy database, GTDB)物种分类中，3 053 个 SGBs 为细菌，可归类为 28 门 782 属；42 个 SGBs 为古菌，可归类为 2 门 8 属。基于 eggNOG 软件功能预测，SGBs 在蛋白相邻类的聚簇(cluster of orthologous groups of proteins, COG)功能分类中可注释到 26 种分类；在京都基因与基因组百科全书(Kyoto encyclopedia of genes and genomes, KEGG)功能预测中，前 25 个直系同源物(KEGG orthology, KO)通路号可归类为 14 种通路类型；碳水化合物酶(carbohydrate-active enzymes, CAZy)预测中，593 个 SGBs 可注释到 6 类碳水化合物酶，分别是辅助氧化还原酶类(auxiliary activities, AA)、碳水化合物酯酶(carbohydrate esterases, CE)、糖苷转移酶(glycosyltransferases, GT)、碳水化合物结合模块(carbohydrate-binding modules, CBM)、糖苷水解酶(glycoside hydrolases, GH)、多糖裂解酶(polysaccharide lyases, PL)；其中，GH 是最为广泛的碳水化合物酶种类。3 095 个 SGBs 加入 Kraken2 标准数据库(2024 年 5 月)后，使得数据库中物种数量增加了 5.00%，数据库大小从 87.2 Gb 提升为 98.2 Gb。通过对一项基于宏基因组技术解析日粮精粗比对荷斯坦奶牛瘤胃微生物组成影

[#]These authors contributed equally to this work.
*Corresponding author. E-mail: hrwang@yzu.edu.cn
Received: 2024-08-03; Accepted: 2024-10-16

响的研究再评估，加入 SGBs 的数据库使得该研究中瘤胃液宏基因组 reads 水平的物种比对率从(19.35±1.81)%提升到(51.04±2.05)%，种水平主成分(principal components analysis, PCA)分析结果表明，扩展的数据库增强了区分 2 种不同日粮精粗比水平下的瘤胃微生物结构的能力，线性判别丰度差异分析(linear discriminant analysis effect size, LEfSe)结果表明，在标准数据库中，*Xylanibacter ruminicola* 和 *Aristaeella hokkaidonensis* 分别是低粗料和高粗料日粮条件下的微生物标志物；而在扩展后的数据库中，*Prevotella* sp. 902800365 和 *Prevotella* sp. 900316445 分别是低粗料和高粗料日粮条件下的微生物标志物。【结论】通过引入 SGBs 扩展 Kraken2 标准数据库，可进一步增加数据库中物种覆盖度，提高宏基因组 reads 水平物种比对率，从而增进对宏基因数据中微生物的理解。

关键词：Kraken2；反刍动物；消化道微生物；分类数据库

Classification ability of extended Kraken2 standard database for digestive tract microbiota in ruminants

WENG Yunan[#], ZHEN Yongkang[#], WANG Mengzhi, WANG Hongrong[*]

College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, Jiangsu, China

Abstract: Metagenomics has enriched our understanding about the composition and functions of digestive tract microbiota in animals. Currently, metagenomic sequencing can generally achieve the classification rate of species between 15% and 45% at the read level. Therefore, improving the alignment rate of microbial reads in metagenomics can help to further mine microbial information from metagenome data. **[Objective]** To enhance the classification ability for digestive tract microbiota in ruminants by extending the Kraken2 standard database, thereby deeply mining the microbial information from metagenome data. **[Methods]** A total of 14 827 metagenome-assembled genomes (MAGs) of the rumen fluid, feces, and digestive tracts of cattle, sheep, and goats were collected. After quality control and filtering, 3 095 species-level genome bins (SGBs) were retained. These SGBs were integrated into the Kraken2 standard database following taxonomic classification and functional prediction, and the classification effect was evaluated. **[Results]** In the genome taxonomy database (GTDB), the 3 095 SGBs were identified as bacteria belonging to 782 genera of 28 phyla (3 053 SGBs) and archaea belonging to 8 genera of 2 phyla (42 SGBs). The functional prediction based on eggNOG annotated the SGBs into 26 clusters of orthologous groups of proteins (COGs). The Kyoto encyclopedia of genes and genomes (KEGG) enrichment categorized the top 25 ortholog groups (KO entries) into 14 pathways. The prediction of carbohydrate-active enzymes (CAZy) showed that 593 SGBs were annotated into six classes of CAZymes: auxiliary activities (AA), carbohydrate esterases (CE), glycosyltransferases (GT), carbohydrate-binding modules (CBM),

glycoside hydrolases (GH), and polysaccharide lyases (PL). Among them, GH was the most common class. The addition of 3 095 SGBs to the Kraken2 standard database (May 2024) increased the number of species in the database by 5.00%, extending the size from 87.2 Gb to 98.2 Gb. Furthermore, a study about the effect of diet fiber-to-concentrate ratio on the rumen microbiota of Holstein cows by metagenomics was reassessed, which showed that the integration of SGBs into the database raised the species alignment rate of rumen metagenome reads from (19.35±1.81)% to (51.04±2.05)%. The principal component analysis results at the species level indicated that the extended database enhanced the ability to distinguish rumen microbiota structures under two different diet fiber-to-concentrate ratios. The linear discriminant analysis effect size results indicated that the microbial markers for low-fiber and high-fiber diets were *Xylanibacter ruminicola* and *Aristaeella hokkaidonensis*, respectively, in the standard database, whereas they were *Prevotella* sp. 902800365 and *Prevotella* sp. 900316445, respectively, in the extended database. **[Conclusion]** In summary, introducing SGBs to extend the Kraken2 standard database can increase species coverage and improve the alignment rate of species at the metagenome read level, thereby enhancing the understanding of microbial information in metagenome data.

Keywords: Kraken2; ruminants; digestive tract microbiota; classification database

反刍动物消化道中栖居的微生物促进了对饲料中营养物质的消化利用[1]。当前的研究表明，消化道中的微生物在动物生命史中对维持宿主健康、改善生产性能、减轻环境负担等方面具有重要作用[2-6]。宏基因组学技术对反刍动物消化道中的微生物功能的表征，拓展了微生物在甲烷减排、胆酸盐代谢、碳水化合物酶谱、抗生素抗性等功能与作用[4,7-10]，并为微生物调控及微生物资源开发利用提供了参考。因而，对宏基因组数据中微生物信息的挖掘，可进一步量化微生物的作用，以及环境、试验处理等因素对微生物的影响。

宏基因组测序读长(reads)水平的微生物物种比对主要通过标记基因、K-mer 以及蛋白质水平 3 种策略进行(表 1)[11]。基于蛋白质水平比对策略的 Kaiju 软件存在运行时间长以及比对结果假阳性高的问题[11]。在当前的研究中，基于标记基因比对策略的 MetaPhlAn 系列软件和基于 K-mer 比对策略的 Kraken2 软件是宏基因 reads 水平物种比对常用的 2 种软件。MetaPhlAn 系列软件基于标记基因的比对算法，具有对未分类物种的注释能力，但纳入数据库的宿主源微生物标记基因的覆盖度会在一定程度上造成比对率低，假阴性高的问题，且对算力需求较大[11]。相比较于 MetaPhlAn 系列软件，Kraken2 系列软件更适合消化道中微生物的分类且更为精准高效[12-14]。Kraken2 主要针对已知物种的比对，计算速度快，但基于标准数据库的宏基因组 reads 水平的物种比对率在 15%−45%[15]，这一局限性导致了宏基因组 reads 水平物种信息挖掘程度不足，进而削弱了我们对微生物作用的理解。因此，引入外源基因组扩展标准数据库物种覆盖范围，提高 Kraken2 软件的物种比对率，可进一步探索宏基因组中微生物信息[15-16]。

表 1 宏基因组 reads 水平比对策略及相应软件比较[11]

Table 1 Metagenome reads level alignment strategy and corresponding software comparison[11]

Alignment strategy	Marker gene		K-mer	Protein
Classifier software	MetaPhlAn		Kraken2	Kaiju
Software version	v3.0	v4.0	v2.1.1	v1.7.4
Third party dependent software	Bowtie2	Bowtie2	None	None
Database version	CHOCOPhlAn 201901	CHOCOPhlAnSGB 202103$^{\Delta}$	plus-pf	nr-euk
Organisms included in the database	Bacteria, archaea, eukaryota	Bacteria, archaea, microbial eukaryotes, virus	Bacteria, archaea, eukaryota, plasmid, human, univec_core, protozoa	Bacteria, archaea, eukaryota, virus, microbial eukaryotes
Database size (Gb)	2.4	23.0	61.0	144.0
Processing time per sample h:m:s$^{\#}$	3:02:38	4:24:14	0:43:29	11:23:26
Classified (%)	5.50	9.40	20.70	62.57

$^{\Delta}$：MetaPhlAn4 的数据库主要包括细菌和古细菌序列，病毒和真核微生物序列的覆盖范围有限；$^{\#}$：CPU 10 核条件下运行时间(h:m:s 为时:分:秒)。

$^{\Delta}$: MetaPhlAn4's database primarily encompasses bacterial and archaeal sequences, with limited coverage of viral and eukaryotic microbial sequences; $^{\#}$: Runtime with a 10-core CPU (h:m:s means hour:minute:second).

当前的外源基因组引入策略中主要依赖引入宿主源的宏基因组组装基因组(metagenome-assembled genomes, MAGs)[7,15-16]。因此，对引入的 MAGs 的亲缘关系以及质量予以评估后过滤，可减少扩展数据库中微生物的冗余。

本研究旨在基于 Kraken2 软件扩展标准数据库对反刍动物消化道微生物的分类能力，并减少在该过程中引入的微生物基因组之间的冗余问题。本研究收集了来自牛、绵羊、山羊瘤胃液、粪便以及消化道中基于宏基因组获得的 MAGs，质控过滤后，获得物种级基因组箱(species-level genome bins, SGBs)，经基因组分类数据库(genome taxonomy database, GTDB)物种分类后，基于获得的物种分类信息，可以扩展 Kraken2 标准数据库(图 1)。以期提高反刍动物消化道宏基因组 reads 水平物种比对率，并为扩展宏基因组分类数据库提供参考。

1 材料与方法

1.1 消化道微生物数据收集

用于消化道微生物构建的数据分别收集于 Stewart、Cao 以及 Zhang 等研究[17-19]，共计 14 827 个 MAGs。

1.2 消化道微生物数据质控

使用 CheckM2 (v1.0.2)软件中的“checkm2 predict”功能用于对收集的 MAGs 的完整性和污染度评估，并使用 fastANI (v1.32)软件中的“fastANI --rl MAGs_genome_list.txt --ql MAGs_genome_list.txt --matrix”功能计算收集的 MAGs 平均核苷酸一致性(average nucleotide identity, ANI)。过滤完整性小于 80%，污染度大于 10%，或 ANI 大于 95%的 MAGs，剩余的 MAGs 认定为 SGBs，并用于 Kraken2 分类数据库扩展。

1.3 SGBs 功能预测

使用 eggNOG (v2.1.12)软件中“emapper.py -i SGBs_genome.fa --itype genome -m diamond”参数对 SGBs 进行功能预测。以描述 SGBs 的蛋白相邻类的聚簇(cluster of orthologous groups of proteins, COG)、直系同源物(KEGG orthology, KO)以及碳水化合物酶(carbohydrate-active enzymes, CAZy)预测信息。

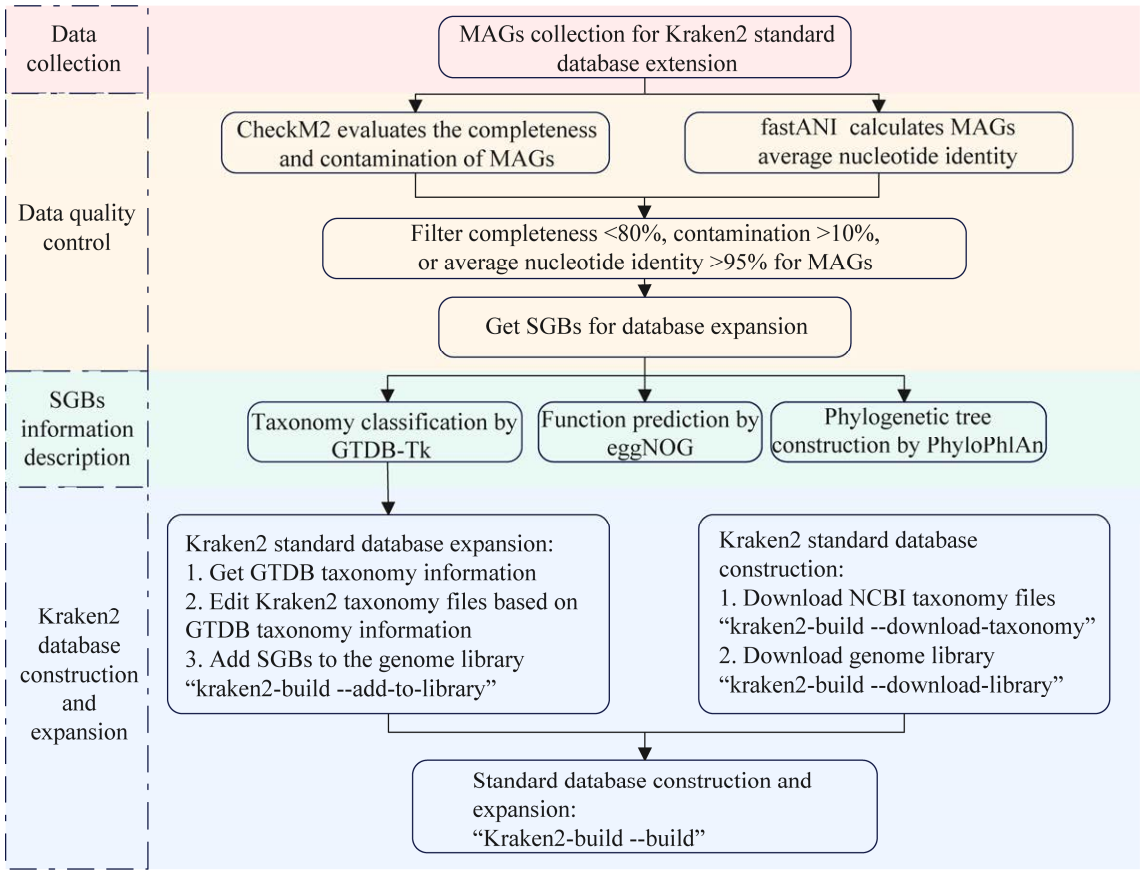


图 1　Kraken2 标准数据库构建及拓展流程图

Figure 1　Construction and expansion process of Kraken2 standard database.

1.4　物种分类以及发育树构建

使用 GTDB-Tk (v2.4.0)软件的“gtdbtk classify_wf”功能，基于 GTDB 分类数据库(220 版本)对 SGBs 进行物种分类，PhyloPhlAn (v3.1.68)软件默认参数用于 SGBs 发育树构建，并使用 iTOL 在线工具(https://itol.embl.de/)对发育树进行可视化。

1.5　标准数据库构建及扩展

基于 Kraken2 软件(v2.1.3)标准数据库构建流程，使用“kraken2-build--download-taxonomy”命令下载物种分类文件，使用“kraken2-build--download-library”命令下载 1 个载体(UniVec_Core)，610 个古菌(archaea)，14 972 个病毒(viral)，109 个真菌(fungi)，46 190 个细菌(bacteria)，2 个人类(human)在内的基因组数据(2024 年 5 月数据)，使用“kraken2-build --build”命令构建标准数据库(RefSeq)。基于获得的 GTDB 分类结果，编辑物种分类文件，并使用“kraken2-build --add-to-library”加入 SGBs，使用“kraken2-build --build”拓展标准数据库(RefSeq+SGBs)。

1.6　奶牛瘤胃宏基因组数据分类效果比较

从 NCBI 数据库中收集试验数据原始测序数据(NCBI BioProject: PRJNA522848)[20]，使用 trimmomatic (v0.39)软件去除原始 reads 中的接头序列，bwa (v0.7.17)软件中的 mem 算法和 samtools (v1.18)软件去除包含苜蓿[21]、玉米(GCA_027171705.3)、人类(GCF_009914755.1 和 GCF_000001405.40)、牛(GCF_002263795.3)、黑麦草(GCA_019359855.2)以及黄豆(GCA_000004515.5)在内的宿主序列，以获得 clean

reads 序列。使用 Kraken2 软件分别基于 RefSeq 数据库和 RefSeq+SGBs 数据库进行比对。

1.7 统计分析

SGBs 基本信息使用“平均值±标准差”描述，并使用 GraphPad Prism (version 9.0.0)进行可视化。R 软件(v4.3.3)用于奶牛宏基因组数据比对结果可视化以及统计分析，结果可视化使用 “ggplot2” 包进行，“geom_pwc(method=“t_test”, label=“p.signif”, p.adjust.method=“fdr”)”功能用于数据统计分析，****表示 P<0.000 1；“factoextra”包用于主成分(principal components analysis, PCA)分析，“microeco”包用于线性判别丰度差异分析(linear discriminant analysis effect size, LEfSe)。

2 结果与分析

2.1 MAGs 质控结果

本研究共收集了来自牛、绵羊和山羊瘤胃液、粪便以及消化道中共计 14 827 个消化道 MAGs，使用 CheckM2 和 fastANI 对收集的 MAGs 的完整度、污染度以及平均核苷酸一致性进行评估，经质控过滤后，保留的 3 095 个 MAGs 被认定为 SGBs，用于数据库构建。SGBs 基本信息如图 2 所示。SGBs 完整度为(90.05±5.78)%，污染度为(1.87±1.92)%，平均基因长度为(328.32±27.28) bp，N50 为(42 589.76±48 639.09) bp，平均基因组大小为(2 179 499.00±738 921.41) bp。

2.2 SGBs 物种分类

使用 GTDB-Tk 软件对 SGBs 进行物种分类，SGBs 的物种组成如图 3 所示，3 053 个 SGBs 被分类为细菌，可归类为 28 门 782 属，在细菌门水平中，*Bacillota_A* 是最丰富的门，占比为 51.29%，在细菌属水平中，*Prevotella* 是最丰富的属，占比为 3.54%；42 个 SGBs 为古菌，可归类为 2 个门，

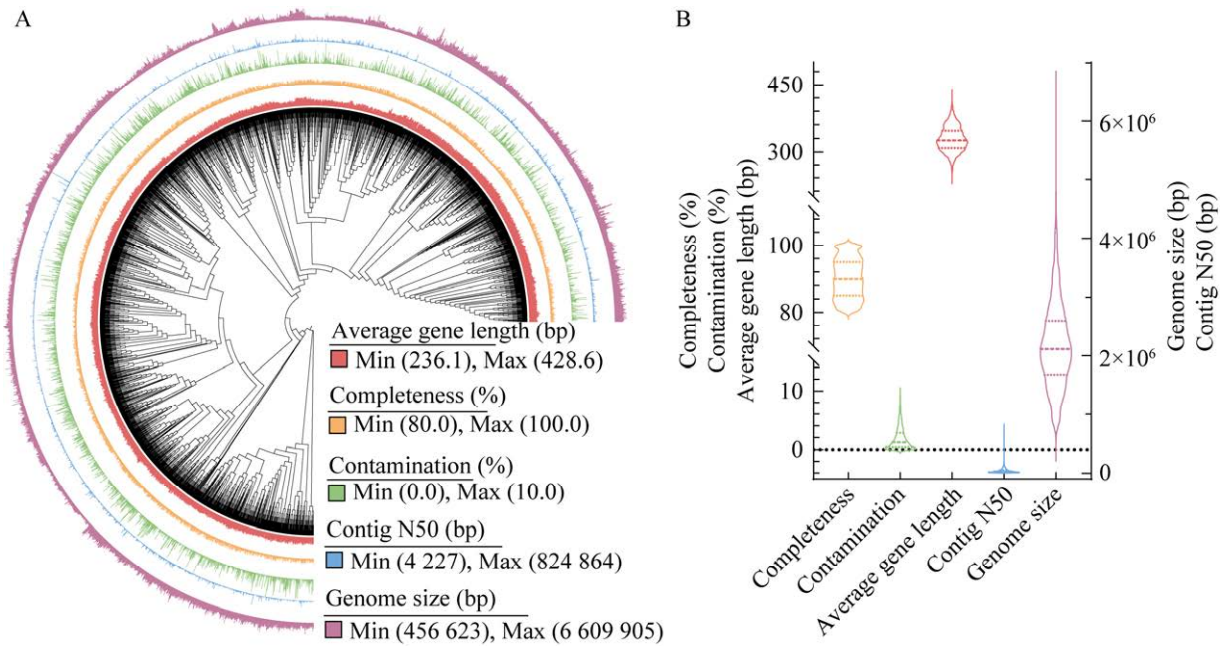


图 2 用于数据库构建的物种级基因组箱(species-level genome bins, SGBs)基本信息

Figure 2 Information of SGBs used for database construction. A: An overview of the average gene length, integrity, contamination, N50, and average genome size distribution of 3 095 SGBs; B: Basic information description for 3 095 SGBs.

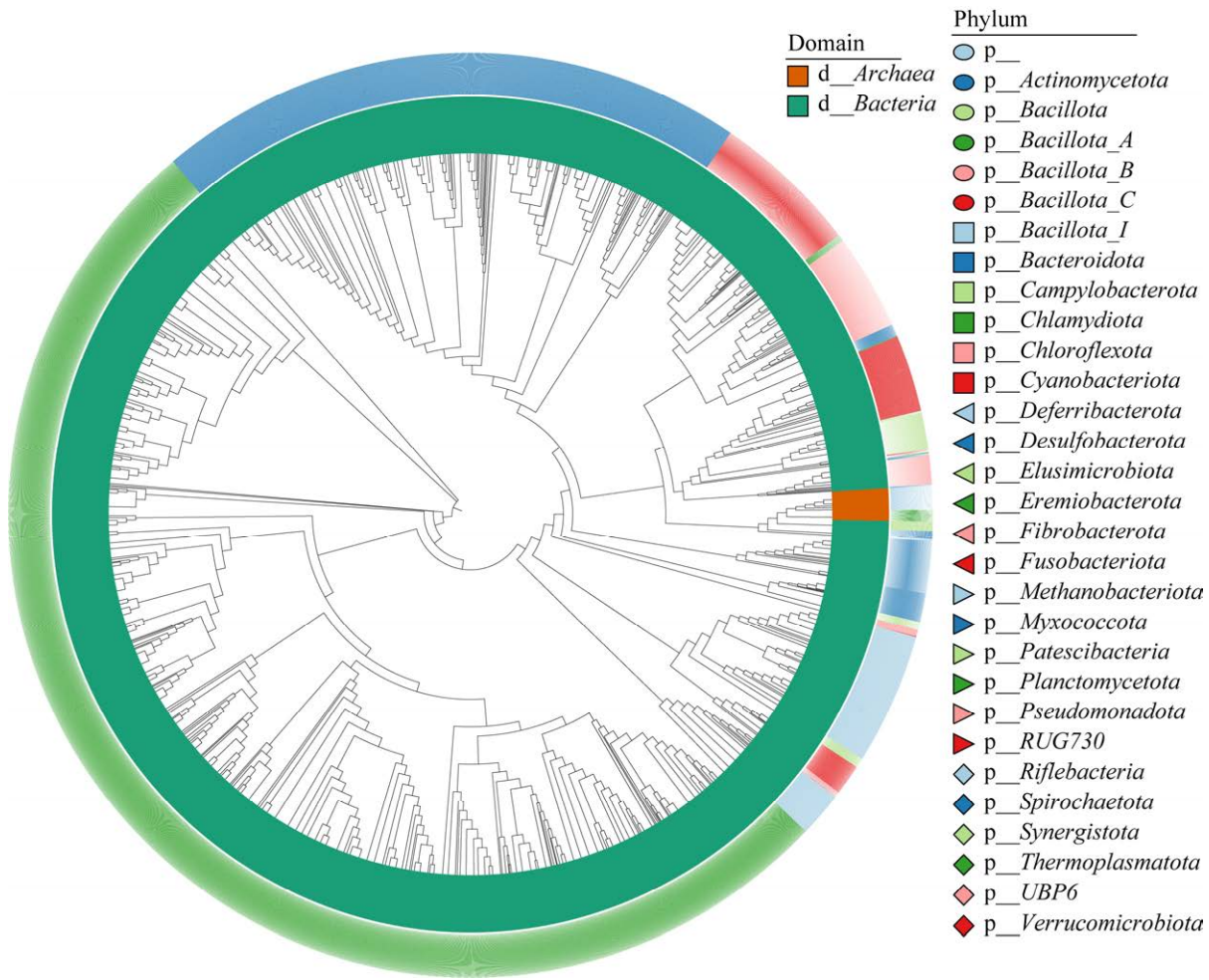


图 3　物种级基因组箱物种分类结果
Figure 3　Taxonomy results of SGBs.

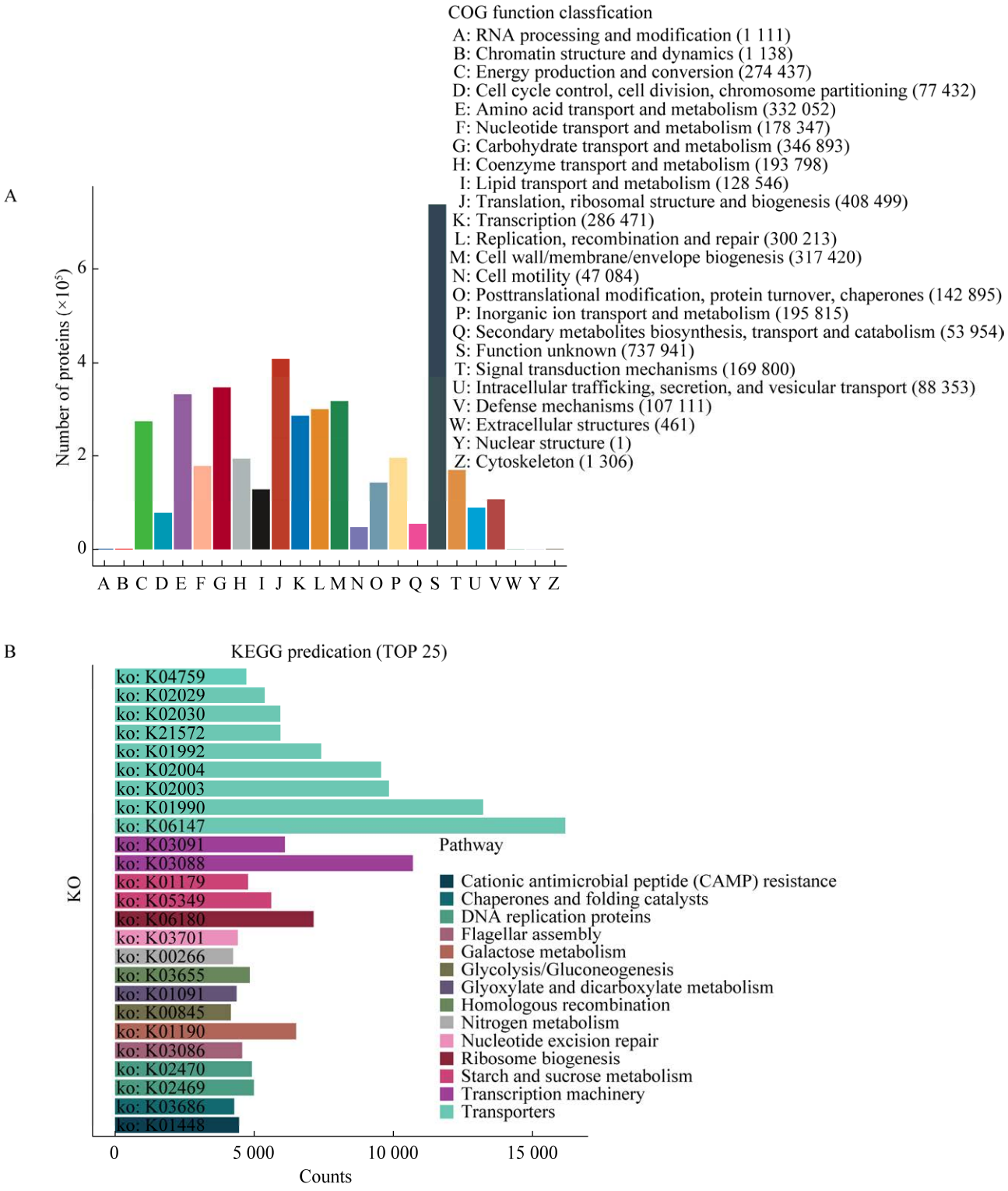
8 个属；在古菌门水平中，*Methanobacteriota* 占比为 59.52%，*Thermoplasmatota* 占比为 40.48%；在古菌属水平中，*Methanobrevibacter_A* 是古菌中最丰富的属，占比为 47.62%。

2.3　SGBs 功能预测分析

SGBs 的 COG 分类、KO 通路以及 CAZy 注释预测结果如图 4 所示。在 COG 分类中，共注释了 26 个 COG 分类。其中，注释最多的 COG 分类是“未知功能”，SGBs 中共有 737 941 个假定基因注释到该分类；注释最少的 COG 分类是“核酸结构”，共有 1 个假定基因注释到该分类(图 4A)。在 KO 通路中，在前 25 个 KO 通路号中，可归类为 14 种 KEGG 通路，共有 9 个 KO 通路与“转运功能”相关(图 4B)。在 CAZy 中，共有 593 个 SGBs 注释到了 97 种碳水化合物酶，可分为辅助氧化还原酶类(auxiliary activities, AA)、碳水化合物酯酶(carbohydrate esterases, CE)、糖苷转移酶(glycosyltransferases, GT)、碳水化合物结合模块(carbohydrate-binding modules, CBM)、糖苷水解酶(glycoside hydrolases, GH)、多糖裂解酶(polysaccharide lyases, PL) 6 类碳水化合物酶类型，其中 GH 是注释最为广泛的碳水化合物类型(图 4C)。

2.4 Kraken2 数据库构建

根据 GTDB 物种分类结果，将 SGBs 加入标准数据库，以构建 RefSeq+SGBs 分类数据库。RefSeq 数据库包含 61 884 个物种基因组序列，大小为 87.2 Gb；RefSeq+SGBs 数据库包含 64 949 个物种基因组序列，大小为 98.2 Gb。相



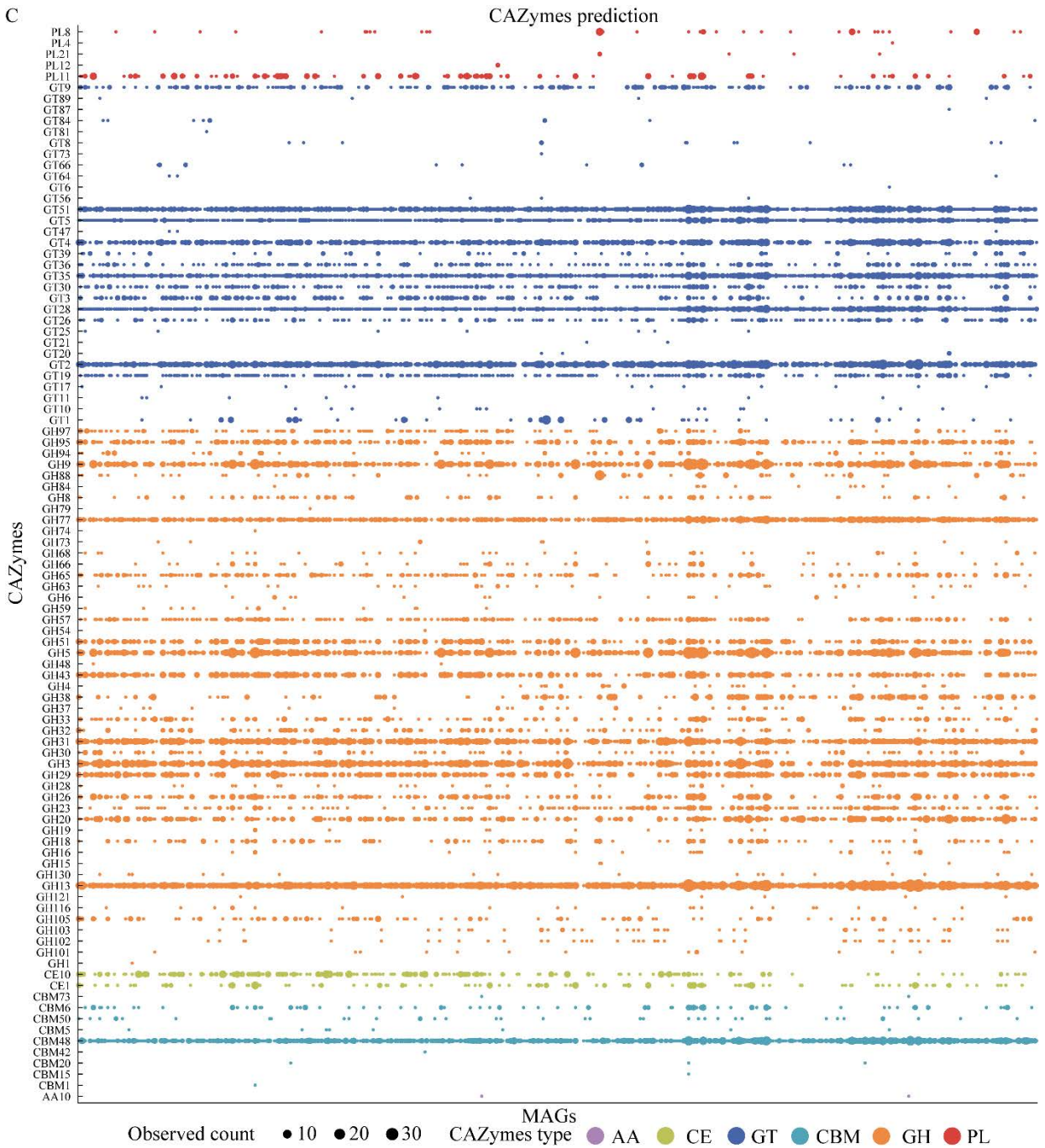


图 4　eggNOG 功能预测结果

Figure 4　Function prediction results of eggNOG. A: The COG function classification result; B: Annotation results for KEGG functionality; C: Predicted results for CAZymes.

较于 RefSeq 数据库，RefSeq+SGBs 数据库中 3 095 个 SGBs 的加入，使得数据库中的物种数量增加了 5.00%，数据库大小增加了 12.61%。

2.5　数据库比对结果比较

用于数据库比对的宏基因数据分别来自饲喂高粗料日粮(HF，精粗比=30:70)和低粗料日粮(LF，精料:粗料=70:30)的荷斯坦奶牛瘤胃液。数据库比对结果如图 5 所示。图 5A 展示了 2 个数据库物种比对率，RefSeq 数据库比对率为(19.35±1.81)%，RefSeq+SGBs 数据库比对率为

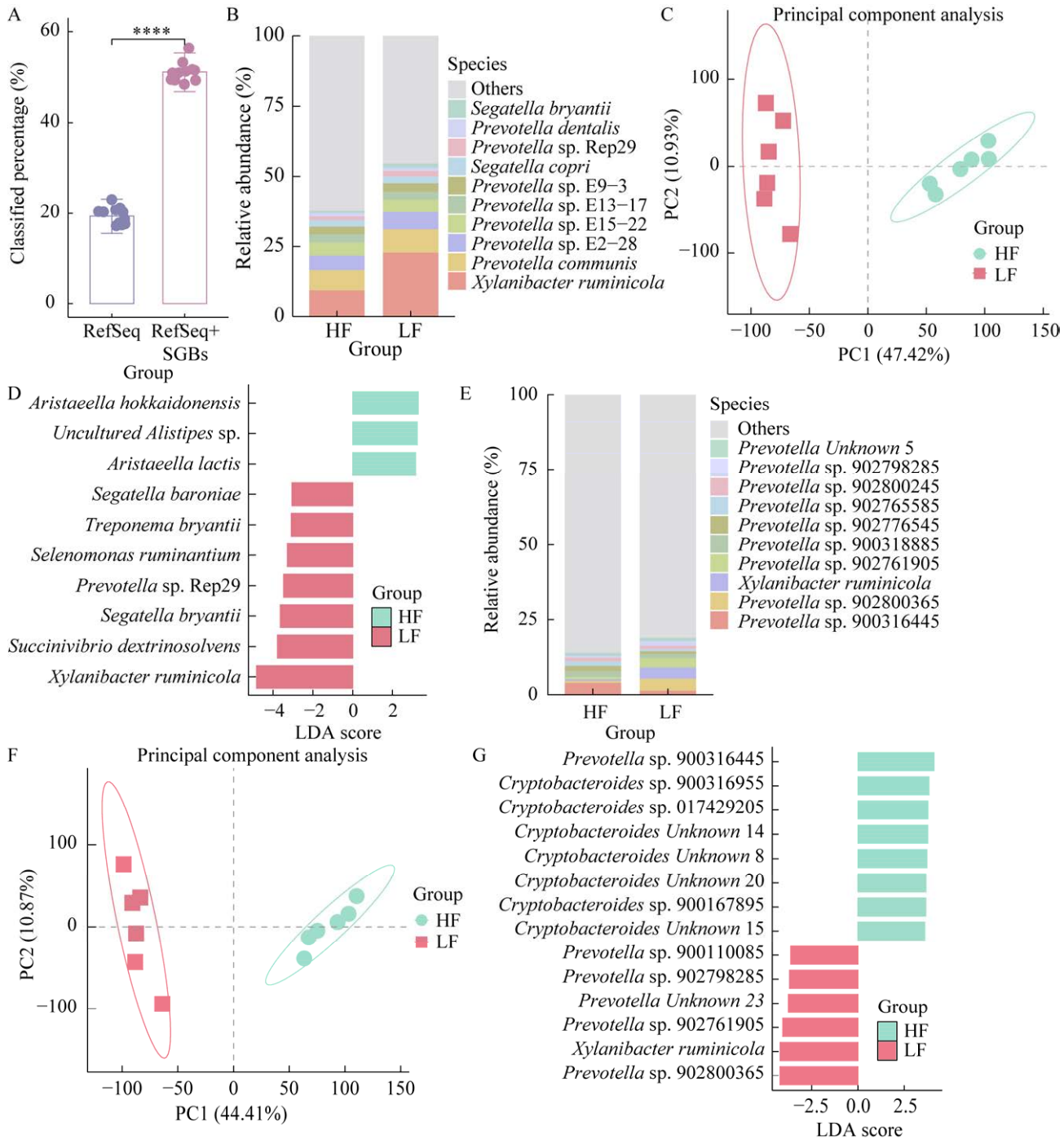


图 5　数据库比对结果比较

Figure 5　Comparison of database alignment results. A: The alignment rate of metagenomic reads, **** means P<0.000 1; B: Based on the species comparison results of RefSeq database at the species level; C: Based on the RefSeq database, the principal component analysis results of the species level comparison were obtained; D: Based on the RefSeq database, the LEfSe analysis results at the species level were obtained; E: Compare species at the species level in the RefSeq+SGBs database; F: Principal component analysis results based on RefSeq+SGBs database level comparison results; G: Based on the RefSeq+SGBs database, the LEfSe analysis results at the species level were obtained.

(51.04±2.05)%，SGBs 的加入显著提高了宏基因组 reads 水平物种比对率(P<0.000 1)；在 RefSeq 数据库比对结果中，相对丰度最高的 10 个物种在 HF 组和 LF 组中分别占 37.67%和 54.69%，*Xylanibacter ruminicola* 和 *Prevotella communis* 是两组中含量最高的 2 个细菌(图 5B)，PCA 结果表明，主成分 1 和主成分 2 分别贡献了 47.42%和 10.93%的解释度(图 5C)，LEfSe 结果显示，*Xylanibacter ruminicola* 是 LF 组 LDA 分数绝对值最高的生物标志物，*Aristaeella hokkaiidonensis* 是 HF 组 LDA 分数绝对值最高的生物标志物(图 5D)；在 RefSeq+SGBs 数据库比对结果中，相对丰度最高的 10 个物种在 HF 组和 LF 组中分别占 14.01%和 18.96%，*Prevotella* sp. 900316445 和 *Prevotella* sp. 902776545 是 HF 组中含量最高的 3 个细菌，*Prevotella* sp. 902800365 和 *Xylanibacter ruminicola* 是 LF 组中含量最高的 2 个细菌(图 5E)，PCA 结果表明，主成分 1 和主成分 2 分别贡献了 44.41%和 10.87%的解释度(图 5F)，LEfse 分析结果表明 *Prevotella* sp. 902800365 是 LF 组 LDA 分数绝对值最高的生物标志物，*Prevotella* sp. 900316445 是 HF 组 LDA 分数绝对值最高的生物标志物(图 5G)。

3 讨论与结论

反刍动物消化道中的微生物与宿主之间的互作，影响了动物的消化功能，以及生产性能、健康状况等诸多生理功能[2-3]。先前基于细菌培养技术获得了微生物的生长参数以及功能特性[22-23]。然而，培养条件和微生物生长特点限制了我们对动物消化道中不可培养微生物的认识，并为这类微生物的体外培养带来了挑战。基因组学的技术手段显著增进了我们对该类微生物的理解[16,24-25]。高通量测序技术的应用在获得微生物遗传信息的同时，可对消化道中不可培养微生物的功能进行表征[24]。相比较于扩增子测序技术，宏基因组的测序深度增加了对物种序列的覆盖度，reads 水平的分析策略可实现对微生物量化观测，因而对序列的识别、归类以及鉴定可进一步了解微生物的组成。

宏基因组 reads 水平的比对可通过标记基因、K-mer 以及蛋白质水平 3 种策略进行(表 1)，Kraken2 基于 K-mer 的比对策略，保证了 Kraken2 在物种分类中的准确性[11,26]。Kraken2 标准数据库基于已知物种微生物构建，缺少对于未分类微生物的鉴定能力，致使宏基因组数据中的比对率在 15%−45%[12,15]。一项对 Kraken2 软件比对能力的研究表明，数据库中物种覆盖度的增加可进一步挖掘宏基因组数据中的微生物信息[27]。引入单一物种源微生物基因组是扩展标准数据库微生物分类能力的可行性方案[7]，在本课题组前期的验证中，4 941 个瘤胃未培养微生物基因组的加入可显著提高物种比对率[17]。MAGs 可表征诸多生物学功能[28]，是数据库引入基因组常见的形式。尽管 MAGs 的引入会在一定程度上增加数据库中物种的覆盖度，但 MAGs 质量以及亲缘关系会在一定程度上造成扩展数据库的冗余。在 Yan 等[29]的研究中，基于 GTDB r207 数据库构建的 Kraken2 数据库中通过引入 MAGs，表明数据冗余对于数据整体比对提升能力有限。在本研究中，过滤了 ANI>95%或完整性<80%，污染度>10%的 MAGs，保留的 MAGs 数量为收集的 MAGs 的 20.87%，保留的 MAGs 被认定为 SGBs，并用以扩展标准数据库，从而均衡多物种源微生物基因组的覆盖度以及数据库中物种冗余。SGBs 物种分类信息的确定可进一步明确宏基因数据比对中物种信息，在本研究中，SGBs 的 GTDB 物种分类结果中，明确了 SGBs 的门、纲、目、

科、属、种信息。在 SGBs 物种分类结果中，本研究中共有 1 个 SGBs 未有明确的门、纲、目、科、属、种分类信息，2 个 SGBs 未有明确的科、属、种分类信息，52 个 SGBs 未有明确的属、种分类信息，1 231 个 SGBs 未有明确的种分类信息，上述未有明确分类信息的 SGBs 被认定为新的微生物物种基因组，据此创建 SGBs 在分类文件中的物种分类信息。未有明确物种分类的 SGBs 在一定程度上增加了对于后续的物种定性的不确定性，对于 MAGs 的过滤应基于更为严格的平均核苷酸一致性、基因组完整度以及污染度参数，从而获得高质量的 SGBs[28-30]，但该方式同时对 MAGs 数据收集以及算力带来挑战。eggNOG 软件对 SGBs 功能预测中，COG 功能分类涵盖了 26 种功能分类，展示了用于数据库扩展的 SGBs 中假定基因功能的多样性，“未知功能”是 COG 中被最多假定基因注释的功能分类，“转运功能”是 KEGG 功能预测中被最多 SGBs 假定基因注释的通路。表明用于数据库扩展的 SGBs 可揭示消化道微生物在物质转运和代谢等方面的多种潜在功能，可为消化道中微生物与宿主之间的通讯、互扰等相互作用提供新见解。此外，19.15%的 SGBs 可预测到碳水化合物酶表达能力，先前的研究表明了瘤胃中宏基因组编码的碳水化合物酶在饲料消化利用上的作用[20,31]，该类 SGBs 的引入可丰富反刍动物消化道微生物对饲料中的碳水化合物消化代谢的认识。

反刍动物瘤胃中微生物会响应日粮的改变，用于数据库验证的宏基因组数据来自饲喂日粮精粗比水平在 30:70 和 70:30 的荷斯坦奶牛瘤胃液。厚壁菌门、拟杆菌门以及变形菌门是瘤胃中最为丰富的菌门，其相对丰度的变化可伴随动物生命史以促进对日粮中营养物质的利用[32]。先前对所收集的宏基因数据的研究中，可表征到日粮类型对碳水化合物酶类丰度的改变[20]。微生物类群精准的碳水化合物酶表达策略增加了对环境中特定碳水化合物多糖的可用性[33]。进一步明确各微生物类群对于碳水化合物酶类的贡献，可增进对微生物参与日粮中碳水化合物代谢的理解。在本研究中，RefSeq+SGBs 数据库的应用使得宏基因数据比对率增加了 163.84%。物种组成中，RefSeq+SGBs 数据库的应用使得在两组中高丰度微生物由 RefSeq 数据库比对结果中木糖杆菌属中的 *Xylanibacter ruminicola* 转变为在 HF 组为普雷沃氏菌属中的 *Prevotella* sp. 900316445 和 LF 组为普雷沃氏菌属中的 *Prevotella* sp. 902800365。PCA 结果表明，标准数据库的扩展提高了瘤胃内未分类微生物作用的权重。LEfSe 分析结果显示，HF 组中的 *Prevotella* sp. 900316445 和 LF 组的 *Prevotella* sp. 902800365 可分别作为两组的生物标志物。当前的研究指出普雷沃氏菌属具有专门处理复杂碳水化合物的基因簇以及多种类型多糖分解能力[34]。本研究对标准数据库的扩展在一定程度上提高了对瘤胃微生物的挖掘程度。然而本研究中，宏基因 reads 水平比对率在 (51.04±2.05)%，对于宏基因 reads 水平的比对率仍有较大的提升空间。因此，对纳入数据库中的 SGBs 应当进一步提高物种覆盖度，同时，还应拓展在不同物种、日粮模式、地理环境、生理状况等条件下收集的 MAGs。对于纳入数据库中的 SGBs 应当满足在物种稀释曲线上的稳定。在 Zhang 等[19]的研究中，纳入数据库的 MAGs 在满足物种稀释曲线稳定的前提下，使得新构建的数据库对宏基因组数据 reads 水平的比对率可达到 79.22%−79.60%。在 Yan 等[29]的研究中，同样可以观察到物种稀释曲线稳定可使得宏基因组数据 reads 水平比对率达到 75%。此外，近年来对于瘤胃内病毒以及原虫的关注[35-38]，

进一步丰富了瘤胃微生态各组分作用；因而，在扩展数据库中还应考虑对病毒、原虫等微生态角色信息的补充。进而增进对消化道微生态的理解。

综上所述，基于 Kraken2 软件，通过引入外源消化道 SGBs，可增加数据库中微生物物种覆盖度，相较于标准数据库，可提高宏基因样品 reads 水平物种比对率，从而增进对反刍动物消化道中微生物的理解。

致谢

衷心感谢扬州大学动物科学与技术学院动物营养与代谢调控研究所在生物信息工作站提供的技术支持。

作者贡献声明

翁玉楠、甄永康：数据收集和处理、论文撰写；王梦芝：项目监管、论文修改；王洪荣：实验设计、论文修改。

作者利益冲突公开声明

作者声明不存在任何可能会影响本文所报告工作的已知经济利益或个人关系。

参考文献

[1] XUE MY, XIE YY, ZANG XW, ZHONG YF, MA XJ, SUN HZ, LIU JX. Deciphering functional groups of rumen microbiome and their underlying potentially causal relationships in shaping host traits[J]. iMeta, 2024, 3(4): e225.

[2] HESS MK, ZETOUNI L, HESS AS, BUDEL J, DODDS KG, HENRY HM, BRAUNING R, McCULLOCH AF, HICKEY SM, JOHNSON PL, ELMES S, WING J, BRYSON B, KNOWLER K, HYNDMAN D, BAIRD H, McRAE KM, JONKER A, JANSSEN PH, McEWAN JC, ROWE SJ. Combining host and rumen metagenome profiling for selection in sheep: prediction of methane, feed efficiency, production, and health traits[J]. Genetics, Selection, Evolution, 2023, 55(1): 53.

[3] XIE F, JIN W, SI HZ, YUAN Y, TAO Y, LIU JH, WANG XX, YANG CJ, LI QS, YAN XT, LIN LM, JIANG Q, ZHANG L, GUO CZ, GREENING C, HELLER R, GUAN LL, POPE PB, TAN ZL, ZHU WY, et al. An integrated gene catalog and over 10, 000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants[J]. Microbiome, 2021, 9(1): 137.

[4] CHAI JM, ZHUANG YM, CUI K, BI YL, ZHANG NF. Metagenomics reveals the temporal dynamics of the rumen resistome and microbiome in goat kids[J]. Microbiome, 2024, 12(1): 14.

[5] DENG J, LIU YJ, WEI WT, HUANG QX, ZHAO LP, LUO LY, ZHU Q, ZHANG L, CHEN Y, REN YL, JIA SG, LIN YL, YANG J, LV FH, ZHANG HP, LI FG, LI L, LI MH. Single-cell transcriptome and metagenome profiling reveals the genetic basis of rumen functions and convergent developmental patterns in ruminants[J]. Genome Research, 2023, 33(10): 1690-1707.

[6] WANG DD, CHEN LY, TANG GF, YU JJ, CHEN J, LI ZJ, CAO YC, LEI XJ, DENG L, WU SR, GUAN LL, YAO JH. Multi-omics revealed the long-term effect of ruminal keystone bacteria and the microbial metabolome on lactation performance in adult dairy goats[J]. Microbiome, 2023, 11(1): 215.

[7] ZHANG BY, JIANG XZ, YU Y, CUI YM, WANG W, LUO HL, STERGIADIS S, WANG B. Rumen microbiome-driven insight into bile acid metabolism and host metabolic regulation[J]. The ISME Journal, 2024, 18(1): wrae098.

[8] GHARECHAHI J, VAHIDI MF, BAHRAM M, HAN JL, DING XZ, SALEKDEH GH. Metagenomic analysis reveals a dynamic microbiome with diversified adaptive functions to utilize high lignocellulosic forages in the cattle rumen[J]. The ISME Journal, 2021, 15(4): 1108-1120.

[9] JING RX, YAN YY. Metagenomic analysis reveals antibiotic resistance genes in the bovine rumen[J]. Microbial Pathogenesis, 2020, 149: 104350.

[10] AUFFRET MD, DEWHURST RJ, DUTHIE CA, ROOKE JA, JOHN WALLACE R, FREEMAN TC, STEWART R, WATSON M, ROEHE R. The rumen microbiome as a reservoir of antimicrobial resistance and pathogenicity genes is directly affected by diet in beef cattle[J]. Microbiome, 2017, 5(1): 159.

[11] EDWIN NR, FITZPATRICK AH, BRENNAN F, ABRAM F, O'SULLIVAN O. An in-depth evaluation of metagenomic classifiers for soil microbiomes[J]. Environmental Microbiome, 2024, 19(1): 19.

[12] TAMAMES J, COBO-SIMÓN M, PUENTE-SÁNCHEZ F. Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes[J]. BMC Genomics, 2019, 20(1): 960.

[13] McINTYRE ABR, OUNIT R, AFSHINNEKOO E, PRILL RJ, HÉNAFF E, ALEXANDER N, MINOT SS, DANKO D, FOOX J, AHSANUDDIN S, TIGHE S, HASAN NA, SUBRAMANIAN P, MOFFAT K, LEVY S, LONARDI S, GREENFIELD N, COLWELL RR, ROSEN GL, MASON CE. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers[J]. Genome Biology, 2017, 18(1): 182.

[14] SZÓSTAK N, SZYMANEK A, HAVRÁNEK J, TOMELA K, RAKOCZY M, SAMELAK-CZAJKA A, SCHMIDT M, FIGLEROWICZ M, MAJTA J, MILANOWSKA-ZABEL K, HANDSCHUH L,

PHILIPS A. The standardisation of the approach to metagenomic human gut analysis: from sample collection to microbiome profiling[J]. Scientific Reports, 2022, 12(1): 8470.
[15] KIESER S, ZDOBNOV EM, TRAJKOVSKI M. Comprehensive mouse microbiota genome catalog reveals major difference to its human counterpart[J]. PLoS Computational Biology, 2022, 18(3): e1009947.
[16] SMITH RH, GLENDINNING L, WALKER AW, WATSON M. Investigating the impact of database choice on the accuracy of metagenomic read classification for the rumen microbiome[J]. Animal Microbiome, 2022, 4(1): 57.
[17] STEWART RD, AUFFRET MD, WARR A, WALKER AW, ROEHE R, WATSON M. Compendium of 4, 941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery[J]. Nature Biotechnology, 2019, 37(8): 953-961.
[18] CAO YH, FENG T, WU YJ, XU YX, DU L, WANG T, LUO YH, WANG Y, LI ZP, XUAN ZY, CHEN SM, YAO N, GAO NL, XIAO Q, HUANG KW, WANG XB, CUI KQ, REHMAN SU, TANG XF, LIU DW, et al. The multi-Kingdom microbiome of the goat gastrointestinal tract[J]. Microbiome, 2023, 11(1): 219.
[19] ZHANG K, HE C, WANG L, SUO LD, GUO MM, GUO JZ, ZHANG T, XU YB, LEI Y, LIU GW, QIAN Q, MAO YR, KALDS P, WU YJ, CUOJI AW, YANG YX, BRUGGER D, GAN SQ, WANG ML, WANG XL, ZHAO FQ, CHEN YL. Compendium of 5 810 genomes of sheep and goat gut microbiomes provides new insights into the glycan and mucin utilization[J]. Microbiome, 2024, 12(1): 104.
[20] WANG LJ, ZHANG GN, XU HJ, XIN HS, ZHANG YG. Metagenomic analyses of microbial and carbohydrate-active enzymes in the rumen of Holstein cows fed different forage-to-concentrate ratios[J]. Frontiers in Microbiology, 2019, 10: 649.
[21] SHEN C, DU HL, CHEN Z, LU HW, ZHU FG, CHEN H, MENG XZ, LIU QW, LIU P, ZHENG LH, LI XX, DONG JL, LIANG CZ, WANG T. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research[J]. Molecular Plant, 2020, 13(9): 1250-1261.
[22] LO GRASSO A, FORT A, MAHDIZADEH FF, MAGNANI A, MOCENNI C. Generalized logistic model of bacterial growth[J]. Mathematical and Computer Modelling of Dynamical Systems, 2023, 29(1): 169-185.
[23] LAGIER JC, EDOUARD S, PAGNIER I, MEDIANNIKOV O, DRANCOURT M, RAOULT D. Current and past strategies for bacterial culture in clinical microbiology[J]. Clinical Microbiology Reviews, 2015, 28(1): 208-236.
[24] BROWNE HP, FORSTER SC, ANONYE BO, KUMAR N, NEVILLE BA, STARES MD, GOULDING D, LAWLEY TD. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation[J]. Nature, 2016, 533(7604): 543-546.
[25] LAGIER JC, DUBOURG G, MILLION M, CADORET F, BILEN M, FENOLLAR F, LEVASSEUR A, ROLAIN JM, FOURNIER PE, RAOULT D. Culturing the human microbiota and culturomics[J]. Nature Reviews Microbiology, 2018, 16: 540-550.
[26] PUSADKAR V, AZAD RK. Benchmarking metagenomic classifiers on simulated ancient and modern metagenomic data[J]. Microorganisms, 2023, 11(10): 2478.
[27] LIU YL, GHAFFARI MH, MA T, TU Y. Impact of database choice and confidence score on the performance of taxonomic classification using Kraken2[J]. aBIOTECH, 2024, 31(7): 1-11.
[28] YOUNGBLUT ND, deLa CUESTA-ZULUAGA J, REISCHER GH, DAUSER S, SCHUSTER N, WALZER C, STALDER G, FARNLEITNER AH, LEY RE. Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity[J]. mSystems, 2020, 5(6): e01045-20.
[29] YAN M, YU ZT. Viruses contribute to microbial diversification in the rumen ecosystem and are associated with certain animal production traits[J]. Microbiome, 2024, 12(1): 82.
[30] MA B, LU CY, WANG YL, YU JW, ZHAO KK, XUE R, REN H, LV XF, PAN RH, ZHANG JB, ZHU YG, XU JM. A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources[J]. Nature Communications, 2023, 14(1): 7318.
[31] BARRETT K, LANGE L, BØRSTING CF, OLIJHOEK DW, LUND P, MEYER AS. Changes in the metagenome-encoded CAZymes of the rumen microbiome are linked to feed-induced reductions in methane emission from Holstein cows[J]. Frontiers in Microbiology, 2022, 13: 855590.
[32] JAMI E, ISRAEL A, KOTSER A, MIZRAHI I. Exploring the bovine rumen bacterial community from birth to adulthood[J]. The ISME Journal, 2013, 7(6): 1069-1079.
[33] GHARECHAHI J, VAHIDI MF, SHARIFI G, ARIAEENEJAD S, DING XZ, HAN JL, SALEKDEH GH. Lignocellulose degradation by rumen bacterial communities: new insights from metagenome analyses[J]. Environmental Research, 2023, 229: 115925.
[34] BETANCUR-MURILLO CL, AGUILAR-MARÍN SB, JOVEL J. *Prevotella*: a key player in ruminal metabolism[J]. Microorganisms, 2022, 11(1): 1.
[35] YAN M, PRATAMA AA, SOMASUNDARAM S, LI ZJ, JIANG Y, SULLIVAN MB, YU ZT. Interrogating the viral dark matter of the rumen ecosystem with a global virome database[J]. Nature Communications, 2023, 14(1): 5254.
[36] SOLOMON R, WEIN T, LEVY B, ESHED S, DROR R, REISS V, ZEHAVI T, FURMAN O, MIZRAHI I, JAMI E. Protozoa populations are ecosystem engineers that shape prokaryotic community structure and function of the rumen microbial ecosystem[J]. The ISME Journal, 2022, 16(4): 1187-1197.
[37] TOYBER I, KUMAR R, JAMI E. Rumen protozoa are a hub for diverse hydrogenotrophic functions[J]. Environmental Microbiology Reports, 2024, 16(4): e13298.
[38] WU YJ, GAO N, SUN CQ, FENG T, LIU QY, CHEN WH. A compendium of ruminant gastrointestinal phage genomes revealed a higher proportion of lytic phages than in any other environments[J]. Microbiome, 2024, 12(1): 69.