



A novel analysis workflow for simultaneous parsing prokaryotic and eukaryotic microbial genes from metagenomes

Wei Zhang^{1,*}, Yanmei Zheng^{2,*}, Guomin Han^{2,3} and Xingbing He⁴

¹ Anhui Province Key Laboratory of Rice Genetics and Breeding (Rice Research Institute, Anhui Academy of Agricultural Sciences), Hefei, Anhui, China

² School of Life Sciences, Anhui Agricultural University, Hefei, Anhui, China

³ National Engineering Laboratory of Crop Stress Resistance Breeding, Anhui Agricultural University, Hefei, Anhui, China

⁴ College of Biology and Environmental Sciences, Jishou University, Jishou, Hunan, China

* These authors contributed equally to this work.

ABSTRACT

Accurately predicting coding genes from metagenomic samples containing a high proportion of eukaryotic content remains a significant challenge. Novel and reliable methods for the simultaneous prediction of prokaryotic and eukaryotic microbial genes are crucial to address this. We evaluated gene prediction accuracy of MetaGeneMark and MetaEuk using representative genomes from diverse organisms. Based on these findings, we developed an innovative analytical workflow. This approach involves an initial prediction of eukaryotic genes using MetaEuk, followed by the masking of these predicted eukaryotic genes and any co-identified partial prokaryotic genes using a custom Perl script. Remaining prokaryotic genes are then predicted from the masked metagenome using MetaGeneMark or metaProdigal. This integrated strategy achieved similar quantities and average lengths of eukaryotic genes compared to using MetaEuk alone. Notably, the quantity of predicted prokaryotic genes and viral genes using the new workflow was 14–18% higher than that obtained with standalone prokaryotic predictors. Furthermore, validation on a mixed prokaryotic-eukaryotic metagenome demonstrated that our workflow yielded genes with significantly higher average lengths, indicating reduced fragmentation and improved gene integrity. This novel workflow effectively enables the rapid and comprehensive retrieval of high-quality prokaryotic and eukaryotic coding sequences from diverse metagenomes.

Subjects Genomics, Microbiology

Keywords Metagenome, Eukaryotic, Prokaryotic, Gene prediction, Novel workflow

INTRODUCTION

The advent of metagenomics, first conceptualized by *Handelsman et al. (1998)*, offered a novel paradigm for studying the collective genomes and biosynthetic machinery of microbial communities, initially focusing on soil microflora. Soil ecosystems harbor a vast diversity of microorganisms, encompassing eubacteria, archaea, and eukaryotes—collectively termed the microbiota—which are integral to processes such as plant growth

Submitted 13 March 2025
Accepted 18 December 2025
Published 11 February 2026

Corresponding authors
Guomin Han,
guominhan@ahau.edu.cn
Xingbing He, hexb@jsu.edu.cn

Academic editor
Rodolfo García-Contreras

Additional Information and
Declarations can be found on
page 9

DOI 10.7717/peerj.20769

© Copyright
2026 Zhang et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

(*Srivastava et al., 2023*). Analytical techniques like marker gene surveys, as well as genomic and metagenomic analyses, have been instrumental in characterizing soil microbiomes and their influencing factors (*Dubey et al., 2020; Knight et al., 2018*). Notably, metagenomic sequencing technology has revolutionized the detection and characterization of microbial species (*Lataretu et al., 2025*), spurring the development of numerous software tools for the taxonomic classification from these complex datasets (*Ye et al., 2019*).

Historically, many widely adopted metagenomic analysis pipelines have primarily concentrated on prokaryotic microorganisms, often leading to an under-representation of eukaryotic components (*Baehren et al., 2023; Dubay et al., 2023*). This focus likely stemmed from the numerical dominance of bacteria in well-studied environments like the human gut and many soil habitats. However, eukaryotic microorganisms are also abundant, particularly in surface soils, where they play crucial roles in nutrient cycling and overall topsoil ecosystem function (*Wu et al., 2022*). Consequently, the analysis of eukaryotic microbial information within metagenomic sequences is essential for a comprehensive understanding of their ecological roles (*Königer et al., 2023*). Encouragingly, software packages capable of analyzing eukaryotic diversity are emerging (*Garfias-Gallegos et al., 2022; Liu et al., 2021*). In the realm of gene prediction from metagenomes, existing tools are typically specialized for either prokaryotic or eukaryotic gene identification. For instance, MetaGeneMark and metaProdigal are widely utilized for predicting prokaryotic genes (*Hyatt et al., 2012; Zhu, Lomsadze & Borodovsky Nar, 2010*), whereas MetaEuk is designed for mining eukaryotic microbial protein-coding genes from metagenomes (*Levy Karin, Mirdita & Söding, 2020*). A significant limitation, however, is that these tools generally do not accurately predict genes across both domains within a single analysis. While EukRep attempts to differentiate eukaryotic and prokaryotic contigs in metagenomics (*West et al., 2018*), its utility can be constrained by issues such as the misassembly of chimeric eukaryotic-prokaryotic contigs (*Pronk & Medema, 2022*).

To address the challenge of comprehensive gene prediction across domains, we have developed an integrated workflow capable of simultaneously identifying both prokaryotic and eukaryotic genes from metagenomic data. Our approach involved an initial evaluation of MetaGeneMark and MetaEuk for gene prediction accuracy across diverse organismal types. Based on these findings, we constructed a novel analytical workflow. This workflow prioritizes the prediction of eukaryotic genes using MetaEuk. Subsequently, these predicted eukaryotic genes, along with any partially predicted prokaryotic genes identified by MetaEuk, are masked within the metagenome. The remaining, unannotated sequences in the masked metagenome are then processed using MetaGeneMark or metaProdigal to predict the remaining prokaryotic genes. This staged, integrated strategy enables a more exhaustive recovery of both prokaryotic and eukaryotic microbial coding sequences from complex metagenomic datasets.

METHODS

To assess the gene prediction accuracy of the prokaryotic pipeline MetaGeneMark and the eukaryotic gene prediction pipeline MetaEuk, we analyzed diverse single-species genomes.

Table 1 Genome and protein sequence information of each species.

Superkingdom	Species	Number of proteins
Archae	<i>Anaerobic archaeon</i>	3,944
	<i>Haloterrigena turkmenica</i>	5,113
	<i>Methanosarcina horonobensis</i>	4,095
	<i>Natrialba magadii</i>	4,204
	<i>Salinigranum rubrum</i>	4,474
Bacteria	<i>Acinetobacter baumannii</i>	3,657
	<i>Bacillus thuringiensis</i>	6,508
	<i>Escherichia coli</i>	5,093
	<i>Vibrio xuii</i>	7,324
	<i>Yersinia pseudotuberculosis</i>	4,340
Fungi	<i>Armillaria gallica</i>	25,695
	<i>Fusarium graminearum</i>	14,145
	<i>Penicillium antarcticum</i>	11,514
	<i>Trametes versicolor</i>	14,296
	<i>Trichoderma asperellum</i>	12,557
Metazoa	<i>Helobdella robusta</i>	23,432
	<i>Lottia gigantea</i>	23,340
	<i>Onchocerca volvulus</i>	14,628
	<i>Pristionchus pacificus</i>	29,202
	<i>Zootermopsis nevadensis</i>	14,610
Plant	<i>Arabidopsis thaliana</i>	48,322
	<i>Brassica rapa</i>	41,025
	<i>Glycine max</i>	88,412
	<i>Oryza sativa</i>	42,411
	<i>Zea mays</i>	131,585

Notes.

The table summarizes protein sequence data for 25 species across five superkingdoms (Archaea, Bacteria, Fungi, Metazoa, and Plant), with five representative species per superkingdom. Protein counts reflect annotated sequences for each species.

For this evaluation, five representative genomes and their corresponding annotated protein sequences were selected for each of the following superkingdoms/kingdoms: Archaea, Bacteria, Fungi, Metazoa, and Plant, all downloaded from NCBI (Table 1).

Based on the evaluation results, we developed a novel analytical workflow for the simultaneous prediction of prokaryotic and eukaryotic genes from metagenomic data (Fig. 1). The workflow proceeds as follows: (1) the assembled metagenomic sequence is first processed with MetaEuk. This step primarily aims to identify eukaryotic organism genes, though some prokaryotic genes may also be detected. (2) Subsequently, a custom Perl script is used to mask the genomic regions corresponding to all genes predicted in this initial step by replacing their sequences with ‘N’ characters (unknown bases). (3) The resulting masked sequences are then subjected to a second round of gene prediction using either MetaGeneMark or metaProdigal to identify remaining prokaryotic genes. (4) Finally, the gene sets predicted by MetaEuk (first pass) and MetaGeneMark/metaProdigal (second

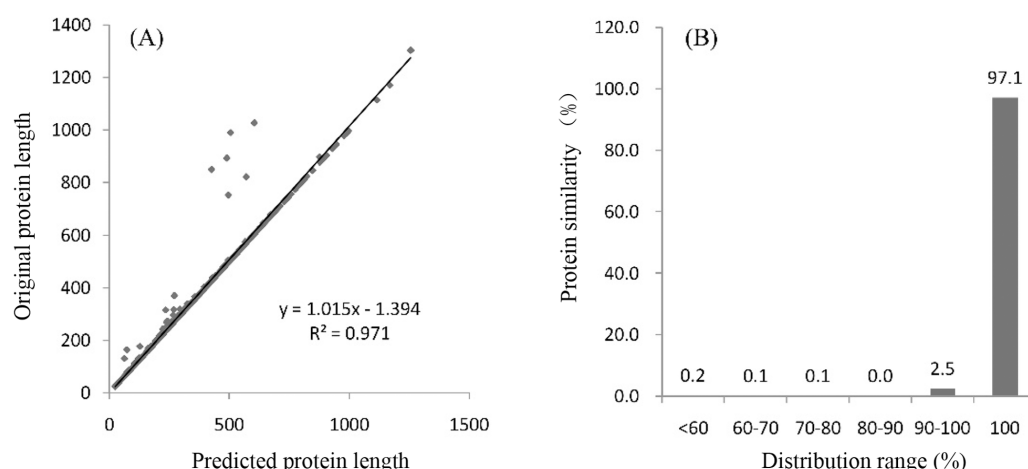


Figure 1 Evaluation results of MetaEuk for prokaryotic organisms. (A) A comparison of the gene length; (B) a comparison of the protein similarity. Note: for a small number of prokaryotic genes predicted by MetaEuk in the prokaryotic prediction results, represented by the results of an archaea, eggNOG was used to evaluate the results. It can be seen from the figure that the prokaryotic results predicted by MetaEuk are very accurate, with 97.1% similarity to the reference sequence proteins.

Full-size [DOI: 10.7717/peerj.20769/fig-1](https://doi.org/10.7717/peerj.20769/fig-1)

pass) are combined to produce a comprehensive set of prokaryotic and eukaryotic gene predictions from the metagenome.

Detailed instructions for implementing this workflow are available on GitHub (<https://github.com/wzhangahas/PPEMS>). For evaluation, we utilized the benchmark metagenomic dataset from the Tara Oceans project, obtained from the MetaEuk repository (<https://wwwuser.gwdguser.de/~compbiol/metaeuk/>). This dataset consists of assembled contigs from marine environments, rich in eukaryotic sequences, and served to test our workflow's performance on complex, eukaryotic-predominant metagenomes. This eukaryote-rich dataset provides a stringent test for eukaryotic gene prediction component of our workflow. The final consolidated set of predicted coding sequences was subjected to functional annotation and orthology assignment using EggNOG-mapper (version 5.0).

In addition, we further validated our workflow using the metagenomic dataset provided by DeepMicroClass (Hou et al., 2024), which contains mixed prokaryotic and eukaryotic contigs. The dataset was assembled by MEGAHIT (v1.2.9). The assembled contigs were analyzed using our integrated workflow, as well as standalone MetaEuk and Metagenemark predictors, followed by functional annotation using EggNOG-mapper.

RESULTS

Comparative evaluation of MetaGeneMark and MetaEuk for gene prediction in diverse single species genomes

We utilized MetaGeneMark and MetaEuk to predict genes from the genomes of diverse single species. The results showed that MetaGeneMark successfully predicted both the number and average length of genes for prokaryotes, closely matching the values obtained from the original genome annotations (Fig. S1). Conversely, according to the results

obtained from MetaGeneMark when applied to eukaryotic genomes, eukaryotic genes were frequently fragmented, resulting in a higher number of genes and shorter average gene lengths compared to the original values (Fig. S1). These results indicate that MetaGeneMark tends to predict a significantly higher number of spurious or fragmented gene segments in eukaryotes, rendering it unsuitable for eukaryotic gene prediction. It is, therefore, primarily optimized for prokaryotic gene prediction.

The MetaEuk software utilizes a homology-based approach to predict coding genes from genomes. MetaEuk predicted a similar number of eukaryotic genes and achieved comparable average gene lengths to the original values (Fig. S1). These findings suggest that MetaEuk is better suited for eukaryotic organisms and provides more reliable gene models.

The MetaEuk software is also capable of predicting a subset of prokaryotic genes. However, the predicted number of prokaryotic genes was significantly lower compared to the original gene count, accounting for only approximately 25% of the original count. Nevertheless, the average lengths of the predicted prokaryotic genes were very close to the original values. Through homology-based comparisons, it was determined that the predictions made by MetaEuk for prokaryotic organisms are highly accurate, with an average sequence identity of 97.1% to reference protein sequences (Fig. 2). Therefore, while MetaEuk primarily focuses on predicting eukaryotic genes in metagenomic data, the subset of prokaryotic genes it predicts is relatively accurate. Although the predicted gene count is significantly lower, these predicted prokaryotic coding genes can be directly used for subsequent analysis.

A workflow for simultaneously mining prokaryotic and eukaryotic microbial genes

Given that MetaEuk is proficient at predicting eukaryotic genes and a minor subset of prokaryotic genes, whereas MetaGeneMark is optimized for prokaryotic gene prediction and performs poorly on eukaryotic sequences (often leading to their fragmentation), a strategic order of application is crucial. Initiating gene prediction with MetaGeneMark would necessitate subsequent complex and computationally intensive steps to distinguish fragmented eukaryotic sequences from true prokaryotic genes. Conversely, employing MetaEuk first yields reliable predictions for the majority of eukaryotic genes and a small, accurate subset of prokaryotic genes, suitable for direct use. Therefore, our novel analytical workflow (Fig. 1) was designed as follows: (1) MetaEuk is used to predict eukaryotic coding genes and a limited number of prokaryotic coding genes from the metagenomic data. (2) These initially predicted genes are then masked. (3) Prokaryotic gene prediction software, such as MetaGeneMark or metaProdigal, is subsequently applied to the masked metagenomic data to identify the remaining prokaryotic coding genes.

A key component of our workflow is an efficient masking step. After MetaEuk identifies potential eukaryotic and a small subset of prokaryotic genes, these sequences must be masked to prevent reprediction by subsequent prokaryotic gene finders. Our initial attempts to use RepeatMasker for this purpose proved impractically slow for large metagenomic datasets; masking an approximately 1 GB dataset took several days, and a 10 GB dataset

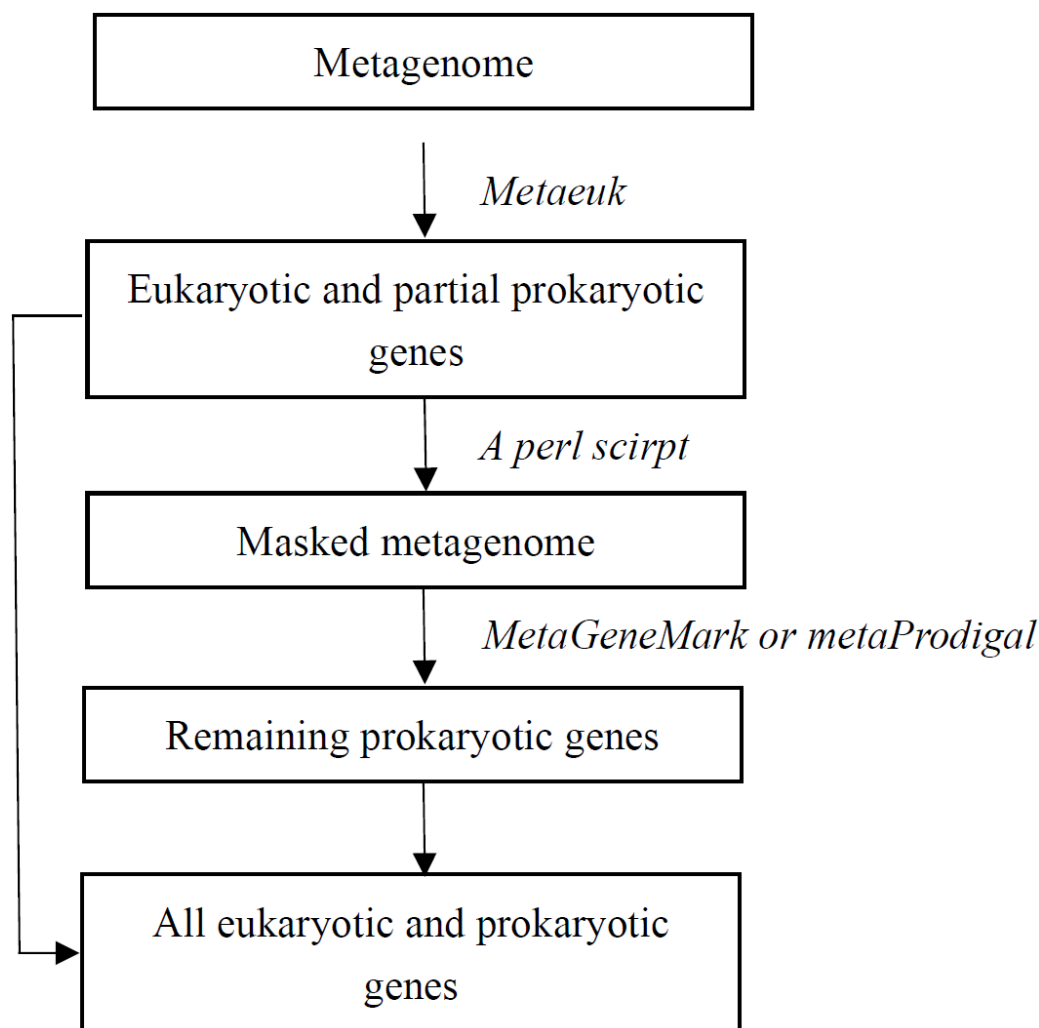


Figure 2 Workflow for simultaneously mining prokaryotic and eukaryotic microbial genes. Note: A workflow that can simultaneously analyse prokaryotic microorganisms, viruses and eukaryotic microorganisms. The workflow involves initially predicting eukaryotic genes using MetaEuk, followed by masking of predicted eukaryotic and partial prokaryotic genes *via* Custom Perl script. The subsequent prediction of remaining prokaryotic genes is conducted using MetaGeneMark or MetaProdigal.

[Full-size](#) DOI: [10.7717/peerj.20769/fig-2](https://doi.org/10.7717/peerj.20769/fig-2)

failed to complete within a month. Recognizing that MetaEuk's output files contain precise gene location information, we developed a custom Perl script. This script directly parses these locations and replaces the corresponding coding sequences in the metagenomic dataset with 'N' characters. This targeted, alignment-free approach reduces the masking time from days or weeks to minutes or hours, significantly improving scalability. Finally, the remaining prokaryotic genes are predicted from this efficiently masked metagenomic sequence using MetaGeneMark or metaProdigal.

The gene predictions from MetaEuk (first pass) and MetaGeneMark/metaProdigal (second pass on masked data) are then consolidated. This final set of predicted coding sequences, encompassing both prokaryotes and eukaryotes, can be subjected to further

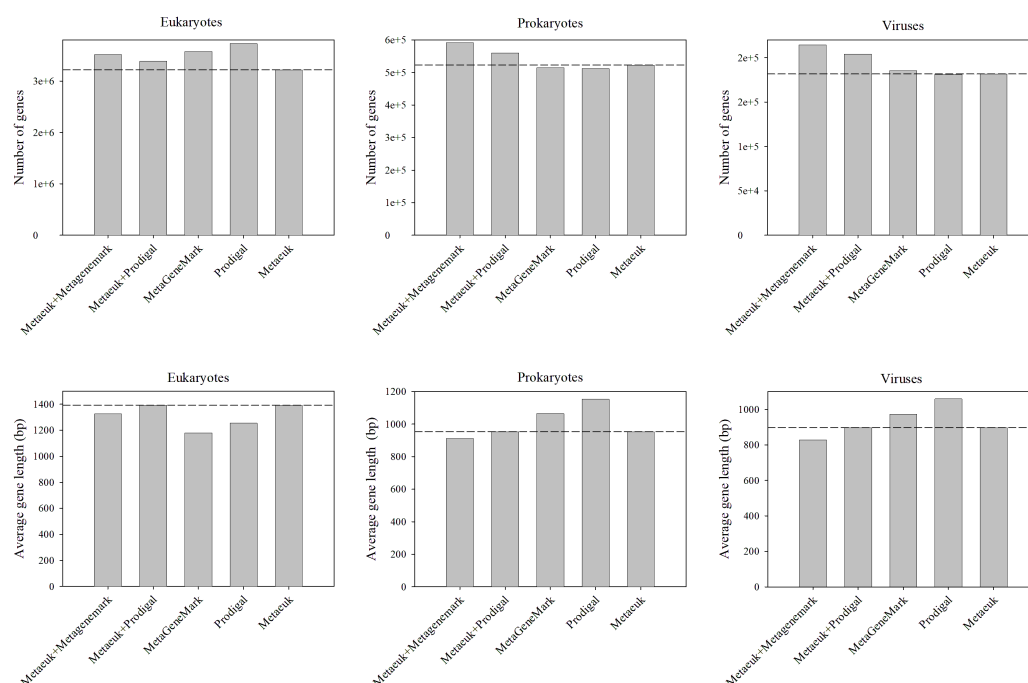


Figure 3 Comparing the new workflow with standalone prokaryotic gene prediction software using real data. To examine this workflow, we analyzed the data in MetaEuk software using a number of different analytical methods, including MetaEuk, MetaGeneMark, and Prodigal. The results showed the number and length of genes calculated by different analysis strategies.

Full-size [DOI: 10.7717/peerj.20769/fig-3](https://doi.org/10.7717/peerj.20769/fig-3)

functional annotation using tools like EggNOG to obtain diverse information such as species origin, Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Clusters of Orthologous Groups of proteins (COG) categories, gene names, and orthologous group assignments.

To evaluate our workflow, we utilized a metagenomic dataset known for its high eukaryotic content (Metagenomic Dataset from Tara Oceans). We compared the performance of our integrated workflow against standalone applications of MetaGeneMark/metaProdigal and MetaEuk (Fig. 3). For eukaryotic gene prediction, our workflow yielded a quantity and average length of genes closely comparable to those obtained by using MetaEuk alone (Fig. 3). In contrast, when compared to standalone prokaryotic predictors (MetaGeneMark/metaProdigal), our workflow identified fewer eukaryotic genes, but these were, on average, considerably longer, suggesting less fragmentation. Regarding prokaryotic genes and viral genes, our workflow predicted 14–18% more genes than standalone prokaryotic predictors. However, the average length of these additional prokaryotic/viral genes was 18–27% shorter.

To further assess the workflow's performance on mixed metagenomes, we analyzed the DeepMicroClass dataset. The results (Fig. S2) demonstrated that our workflow predicted more genes than the standalone methods, although the mean gene length was marginally shorter than that obtained with metaProdigal alone.

DISCUSSION

The rapid advancement of metagenomic sequencing analysis has spurred the development of specialized tools for gene prediction ([Navgire et al., 2022](#)); however, a persistent challenge lies in the comprehensive and accurate annotation of coding sequences across both prokaryotic and eukaryotic domains from mixed microbial community data. Our comparative assessment using representative single-species genomes confirmed distinct performance profiles: while prokaryotic predictors like MetaGeneMark accurately identify prokaryotic genes, they tend to erroneously fragment eukaryotic genes. Conversely, the eukaryotic gene predictor MetaEuk excels at identifying eukaryotic genes with high fidelity and also captures a smaller, yet reliable, subset of prokaryotic coding genes. These observations highlight the inadequacy of standalone tools for robust, cross-domain gene prediction in complex metagenomes.

The primary difficulty in analyzing metagenomes with substantial eukaryotic content is the accurate and efficient prediction of complete coding genes, particularly intron-containing eukaryotic genes, alongside prokaryotic genes. Our study demonstrates that a strategically designed, staged workflow effectively addresses this. By prioritizing the initial identification of eukaryotic genes using MetaEuk, our workflow preserves the integrity of these sequences, thereby avoiding the extensive fragmentation typically observed when prokaryotic predictors are directly applied to eukaryotic DNA. While MetaEuk itself identifies some prokaryotic and viral genes, our integrated approach further enhances the recovery of these genes by approximately 14–18% through a subsequent targeted prediction using MetaGeneMark or metaProdigal on sequences masked by the initial MetaEuk predictions. Importantly, this workflow ensures that the quantity and average length of predicted eukaryotic genes remain comparable to those obtained using MetaEuk alone, while significantly improving the yield and diversity of prokaryotic and viral genes compared to the application of standalone prokaryotic tools. The observation that these additionally predicted prokaryotic/viral genes have a shorter average length warrants further investigation. This could potentially indicate the recovery of partial or incomplete genes from fragmented contigs, or alternatively, complete genes from organisms that inherently possess shorter coding sequences, which might be overlooked by more stringent standalone prokaryotic predictors ([Li et al., 2021](#)).

Our validation with the DeepMicroClass dataset further confirms the robustness of the workflow. Compared with standalone MetaEuk or Metagenemark, the integrated approach yielded more genes with comparable annotation completeness, indicating enhanced prediction precision for mixed metagenomic environments. Moreover, while tools such as DeepMicroClass and Whokaryote ([Pronk & Medema, 2022](#)) specialize in separating contigs by domain prior to gene prediction, our workflow focuses on gene-level integration across domains after assembly. These approaches are complementary, and future versions of our pipeline could incorporate such classification steps to improve pre-processing efficiency. As sequencing costs continue to decrease and the scale of metagenomic datasets expands, the computational efficiency of each analytical step becomes increasingly critical. Conventional alignment-based masking approaches, such as those employed by tools like RepeatMasker

(often designed for repeat identification rather than gene masking), are prohibitively slow for processing large metagenomic datasets after initial gene prediction. Our custom Perl script circumvents this significant bottleneck. By directly parsing gene coordinates from MetaEuk's output and performing in-memory replacement of coding regions with 'N' characters, this script drastically reduces masking time from days or weeks to mere minutes or hours. This efficiency is a cornerstone of our workflow's practicality and scalability, enabling the rapid and comprehensive retrieval of both eukaryotic and prokaryotic coding sequences.

In conclusion, the integrated workflow presented here offers a pragmatic and efficient solution for the simultaneous prediction of prokaryotic and eukaryotic genes from metagenomic data, particularly from samples rich in eukaryotic content. By leveraging the strengths of domain-specific predictors in a sequential manner, coupled with a highly efficient custom masking procedure, our approach enhances gene recovery while maintaining the integrity of eukaryotic gene structures. The consolidated and more comprehensive gene sets generated by this workflow provide a robust foundation for subsequent downstream functional analyses, such as metabolic pathway reconstruction and phylogenetic studies, thereby facilitating deeper investigations into the functional roles and interactions within complex microbial ecosystems.

CONCLUSIONS

The new workflow effectively enables the rapid and precise retrieval of both prokaryotic and eukaryotic microbial genes from metagenomes.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the National Nature Science Foundation of China (32060332). The Anhui Provincial Key Research and Development Program (Grant No: 2023h11020007) supported the APC for this article. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

National Nature Science Foundation of China: 32060332.

The Anhui Provincial Key Research and Development Program: 2023h11020007.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Wei Zhang conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

- Yanmei Zheng conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Guomin Han conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Xingbing He conceived and designed the experiments, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at MetaEuk: <https://wwwuser.gwdg.de/~compbiol/metaeuk/>.

The code is available at GitHub and Zenodo:

– <https://github.com/wzhangahas/PPEMS>.

– Zhang, W. (2025). Parsing Prokaryotic and Eukaryotic Microbial Genes Simultaneously from Metagenome by a Novel Analysis Workflow. Zenodo. <https://doi.org/10.5281/zenodo.18027880>.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.20769#supplemental-information>.

REFERENCES

- Baehren C, Pembaur A, Weil PP, Wewers N, Schult F, Wirth S, Postberg J, Aydin M. 2023. The overlooked microbiome-considering archaea and eukaryotes using multiplex nanopore-16S-/18S-rDNA-sequencing: a technical report focusing on nasopharyngeal microbiomes. *International Journal of Molecular Sciences* 24(2):1426 DOI 10.3390/ijms24021426.
- Dubay MM, Acres J, Rieckes M, Nadeau JL. 2023. Recent advances in experimental design and data analysis to characterize prokaryotic motility. *Journal of Microbiological Methods* 204:106658 DOI 10.1016/j.mimet.2022.106658.
- Dubey RK, Tripathi V, Prabha R, Chaurasia R, Singh DP, Rao CS, El-Keblawy A, Abhilash PC. 2020. *Unravelling the soil microbiome: perspectives for environmental sustainability*. Cham: Springer DOI 10.1007/978-3-030-15516-2.
- Garfias-Gallegos D, Ziri6n-Mart6nez C, Bustos-D6az ED, Arellano-Fern6ndez TV, Lovaco-Flores JA, Espinosa-Jaime A, Avelar-Rivas JA, S6lem-M6jica N. 2022. Metagenomics bioinformatic pipeline. *Methods in Molecular Biology* 2512:153–179 DOI 10.1007/978-1-0716-2429-6_10.
- Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* 5:R245–R249 DOI 10.1016/S1074-5521(98)90108-9.
- Hou S, Tang T, Cheng S, Liu Y, Xia T, Chen T, Fuhrman Jed A, Sun F. 2024. DeepMicroClass sorts metagenomic contigs into prokaryotes, eukaryotes and viruses. *NAR Genomics and Bioinformatics* 6(2):lqae044 DOI 10.1093/nargab/lqae044.

- Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28:2223–2230 DOI 10.1093/bioinformatics/bts429.
- Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciulek T, McCall L-I, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swofford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. 2018. Best practices for analysing microbiomes. *Nature Reviews Microbiology* 16:410–422 DOI 10.1038/s41579-018-0029-9.
- Köninger J, Ballabio C, Panagos P, Jones A, Schmid MW, Orgiazzi A, Briones MJL. 2023. Ecosystem type drives soil eukaryotic diversity and composition in Europe. *Global Change Biology* 29:5706–5719 DOI 10.1111/gcb.16871.
- Lataretu M, Krautwurst S, Huska MR, Marquet M, Viehweger A, Braun SD, Brandt C, Hölzer M. 2025. Targeted decontamination of sequencing data with CLEAN. *bioRxiv* DOI 10.1101/2023.08.05.55208.
- Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8:48 DOI 10.1186/s40168-020-00808-x.
- Li W, O'Neill KR, Haft DH, Di Cuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire M, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Research* 49(D1):D1020–D1028 DOI 10.1093/nar/gkaa1105.
- Liu YX, Qin Y, Chen T, Lu M, Qian X, Guo X, Bai Y. 2021. A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* 12:315–330 DOI 10.1007/s13238-020-00724-8.
- Navgire S, Goel N, Sawhney G, Sharma M, Kaushik P, Mohanta YK, Mohanta TK, Al-Harras A. 2022. Analysis and interpretation of metagenomics data: an approach. *Biological Procedures* 24:18 DOI 10.1186/s12575-022-00179-7.
- Pronk LJU, Medema MH. 2022. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microbial Genomics* 8(5):mgen000823 DOI 10.1099/mgen.0.000823.
- Srivastava P, Sachan K, Baskar P, Saikanth DRK, Lytand W, Kumar RKM, Singh BV. 2023. Soil microbes expertly balancing nutrient demands and environmental preservation and ensuring the delicate stability of our ecosystems- a review. *International Journal of Plant & Soil Science* 35:989–1000 DOI 10.9734/ijpss/2023/v35i183363.
- West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research* 28:569–580 DOI 10.1101/gr.228429.117.

- Wu JY, Ding FG, Shen ZW, Hua ZL, Gu L. 2022.** Linking microbiomes with per- and poly-fluoroalkyl substances (PFASs) in soil ecosystems: microbial community assembly, stability, and trophic phyllosymbiosis. *Chemosphere* **305**:135403 DOI [10.1016/j.chemosphere.2022.135403](https://doi.org/10.1016/j.chemosphere.2022.135403).
- Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019.** Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**:779–794 DOI [10.1016/j.cell.2019.07.010](https://doi.org/10.1016/j.cell.2019.07.010).
- Zhu W, Lomsadze A, Borodovsky Nar MJ. 2010.** Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* **38**(12):e132 DOI [10.1093/nar/gkq275](https://doi.org/10.1093/nar/gkq275).