BMC Genomics

DATABASE

Open Access

Exploring microproteins from various model organisms using the mip-mining database

Bowen Zhao[1†], Jing Zhao[1†], Muyao Wang[1], Yangfan Guo[2], Aamir Mehmood[1], Weibin Wang[1], Yi Xiong[1,3], Shenggan Luo[1], Dong-Qing Wei[1,4,5*], Xin-Qing Zhao[1*] and Yanjing Wang[1,6*]

Abstract

Microproteins, prevalent across all kingdoms of life, play a crucial role in cell physiology and human health. Although global gene transcription is widely explored and abundantly available, our understanding of microprotein functions using transcriptome data is still limited. To mitigate this problem, we present a database, Mip-mining (https://weilab.sjtu.edu.cn/mipmining/), underpinned by high-quality RNA-sequencing data exclusively aimed at analyzing microprotein functions. The Mip-mining hosts 336 sets of high-quality transcriptome data from 8626 samples and nine representative living organisms, including microorganisms, plants, animals, and humans, in our Mip-mining database. Our database specifically provides a focus on a range of diseases and environmental stress conditions, taking into account chemical, physical, biological, and diseases-related stresses. Comparatively, our platform enables customized analysis by inputting desired data sets with self-determined cutoff values. The practicality of Mip-mining is demonstrated by identifying essential microproteins in different species and revealing the importance of *ATP15* in the acetic acid stress tolerance of budding yeast. We believe that Mip-mining will facilitate a greater understanding and application of microproteins in biotechnology. Moreover, it will be beneficial for designing therapeutic strategies under various biological conditions.

Keywords Microprotein, RNA-Seq, Stress response, Disease, Mip-mining database

Background

Microproteins, also called small proteins, or miniproteins, are encoded by small open reading frames (smORFs). Microproteins generally refer to proteins composed of up to 50 and 100 amino acids in prokaryotes and eukaryotes, respectively [1, 2]. Genes encoding such proteins are commonly presented in almost all domains of life, including bacteria, fungi, insects, plants, animals, and human microbiomes [2, 3]. However, related functional studies have been limited and even neglected, probably due to their small size and difficulty in detection due to low abundance and or special properties[3].

†Bowen Zhao and Jing Zhao contribute equally to this work.

*Correspondence:
Dong-Qing Wei
dqwei@sjtu.edu.cn
Xin-Qing Zhao
xqzhao@sjtu.edu.cn
Yanjing Wang
wangyanjing@sjtu.edu.cn
[1]State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Central Laboratory of Yan'an Hospital Affiliated to Kunming Medical University, Kunming 650051, China
[3]Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[4]Zhongjing Research and Industrialization Institute of Chinese Medicine, Zhongguancun Scientific Park, Meixi, Nayang, Henan 473006, China
[5]Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nanshan District, Shenzhen 518055, Guangdong, China
[6]Engineering Research Center of Cell & Therapeutic Antibody, School of Pharmacy, Shanghai Jiao Tong University, Shanghai 200240, China

Recently, studies on microproteins as 'dark matter' in proteomics have received increasing attention [4]. Various studies have reported discovering and characterizing smORFs and microproteins in different living organisms, including microorganisms, plants, and humans [5–8]. It was has revealed that some microproteins are essential in cellular physiology, metabolism, development, cell signaling, and disease occurrence in various living organisms [9–15]. With the increasingly accumulated data available on the existence and expression of microproteins in multiple organisms, it will be feasible to unveil the functions and working mechanisms of this family of proteins.

Among the known functions of microproteins, cellular stress responses are of particular interest in various fields, including biology, biotechnology, and medical science [2]. Cells are confronted by constant changes in their external environmental conditions. During growth and metabolism, cells may encounter harsh environments, e.g., low pH, oxidative stress, high temperature, and toxins. Studies on microbial stress tolerance have received significant attention due to their implications in cell metabolism, environmental toxicity, food preservation, and fermentation efficiency to produce biofuels and biochemicals [16–20]. For example, the development of stress-tolerant yeast strains benefits efficient fuel ethanol production [21]. For higher eukaryotes such as plants and humans, failure to combat stressful environments leads to developmental deficiency and or diseases [18–20]. Therefore, stress tolerance has been an important topic for the developmental process, breeding crops, and disease treatment.

It has been reported that many microproteins participate in stress response and tolerance [2]. The development of efficient high-throughput gene manipulating methods, for example, CRISPR-based genome editing tools, has enabled rapid characterization of microprotein gene functions [11]. In addition, synthetic biology approaches can be employed to design and manipulate microproteins for improved phenotypes. Therefore, it can be expected that studying microprotein functions in stress response and tolerance substantially impacts microbial biotechnological applications, agriculture, longevity, and human health [22].

So far, there have been multiple databases collecting multi-layered information on microproteins, for example, the plant-related ones, namely, *Arabidopsis thaliana*-oriented microprotein database ARA-PEPs [23]; and plant-oriented microprotein database PsORF [24]; as well as SmProt [25, 26] which is based on eight model organisms (*Escherichia coli*, yeast, zebrafish, rat, mouse, fruit fly, *Caenorhabditis elegans*, and human) integrating multi-source microprotein data mainly in Ribosome profiling sequencing (Ribo-seq) data and mass spectrometry data. In addition, OpenProt [27, 28] was developed for small protein mining based on eukaryotes; TISdb [29] for alternative translation initiation in mammalian cells, and SORFs.org [30, 31], a database of small ORFs using Ribo-seq data. However, there are several limitations of the current databases: (1) The species covered by the databases mentioned above are limited in specific domains of life (mostly plants, microbes and or animals); (2) Most of these databases only provide search results and cannot perform personalized analysis [32–34]; (3) Transcriptomic data have been largely overlooked. Transcription regulation is critical for gene expression, and transcriptome data are abundantly available, which benefits exploring differential transcription of possible microprotein-encoding genes and their related genes for functional characterization [35]. (4) No database has been developed to explore microproteins involved in responses to environmental stress and diseases, which are critical to sustainable bioproduction and disease treatment.

To address the above limitations, we have developed a microprotein mining database called Mip-mining, and made a collection of 336 sets of RNA-seq data from species ranging from *Escherichia coli* to humans. The database presented here is designed explicitly for probing microprotein functions, which enables locating functional microproteins under stress conditions in a particular species or various diseases, especially cancers. Our database benefits the exploration of microprotein functions in stress response and disease occurrence, which are receiving increasing attention in various fields [36, 37]. We also demonstrate the identification of essential microproteins in budding yeast, plants, and humans using Mip-mining.

Construction and content

Database content

A total of 336 sets of data were deposited in the current version of our database covering nine species, including *A. thaliana*, *E. coli*, *Oryza sativa*, *Saccharomyces cerevisiae*, *C. elegans*, *Danio rerio*, *Drosophila melanogaster*, *Mus musculus*, and *Homo sapiens*. Each set of the data contains specific information: the GSE Accession of the RNA-seq data in the GEO database, the stress type of the experiment, the sample number of the data, and the source of the RNA-seq data, including the GSE title with the corresponding link. Each data set has been manually checked and processed using a high-performance computing platform through a standard RNA-seq analysis process. Redundant intermediate files are deleted to save the time of users and computer storage space.

Data collection and organization

To reveal the relationship between microprotein and its function, we chose to collect stress-related data by using

keywords such as "stress" and "response to" to search in the GEO database [38]. Human diseases such as "diabetes" and "cancer" are also related to stress [39], so we also added these data. The data as a whole has been manually checked to ensure that it retains the original data and that it belongs to RNA-seq files. Additionally, the corresponding literature was also checked to confirm whether the results were related to stress. For a dataset to be included in our database, the corresponding relevant dataset was selected to meet the following predefined inclusion criteria: (1) The original SRA file is available; (2) a related study for related research has been published and can be tracked; and (3) enough relevant RNA-seq data is available to construct at least one comparison model. Finally, we categorized the data according to species and stress types. The number of stress types in the database is listed in Table 1.

Reference genome resources and reference microproteins

Each species' reference genome and annotation files were downloaded from GENCODE [40], Ensembl [41], and the NCBI-Genome database. The downloaded files contain reference genome fasta data, index data during Hisat2 [42] alignment, and general feature format (gtf) data.

Reference microproteins were obtained by a two-step screening. First, all microproteins ($\leq$ 100 AA) related to each species were downloaded from the UniProt database (https://www.uniprot.org/). Importantly, considering that most microproteins are large protein fragments or recognizable subunits, we performed a second round of screening and obtained high-confidence reference microproteins.

Expression matrix retrieval

We used the standard RNA-seq procedure to process the selected high-quality transcriptomic data. Sratoolkit (https://github.com/ncbi/sra-tools/wiki) is a toolkit provided by NCBI for processing sequencing data from the SRA database (Sequence Read Archive database [43]), and we used its built-in plugins for data processing.

Table 1 Number of stress conditions per species

Species	Biological response	Chemical treatment	Disease	Multiple stresses	Physical stress
S. cerevisiae	4	37	0	3	6
O. sativa	2	3	0	0	14
E. coli	0	37	0	3	6
A. thaliana	0	6	0	0	22
M. musculus	3	29	0	0	6
H. sapiens	3	17	91	0	9
D. rerio	3	12	0	0	9
C. elegans	6	8	0	0	9
D. melanogaster	1	9	0	0	2

The Prefetch (version 2.10.9) was used to download the data, and fasterq-dump (version 2.10.9) assisted in decompressing data. In terms of quality control, we used FastQC (version 0.11.9) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) to check data quality, multiQC (version 1.9) [44] to integrate data quality files, and fastp (version 0.19.5) [45] to cut low-quality fragments to ensure the quality of each set of information. Next, we aligned the sequencing files to the reference genome using Hisat2 (version 2.2.1) [42], and we employed StringTie (version 2.1.4) [46] to generate merged transcripts, before converting them to the format adapted for the downstream processing R package called ballgown (version 2.18.0) [47].

Differential expression and enrichment

We conducted statistical analyses in the R environment (Version 3.6.1, http://cran.r-project.org/). Several R packages were used; for instance, the "ballgown" (version: 2.18.0) constructed the gene FPKM expression matrix; for principal component analysis, we used "factoextra"(version: 1.0.7) [48] and "FactoMineR" (version: 2.4) [49] packages for data dimensionality reduction. The differentially expressed microproteins were screened using the "limma" package (version: 3.42.0) [50]. Downstream enrichment analysis, including GO, KEGG, and GSEA annotations, are performed through these packages: "enrichplot" (version:1.6.0) [51], and "clusterProfiler" (version:3.14.0) [52]. Visualization of analysis results is achieved by integrating the "ggplot2" package (version: 3.3.3) [53] with the "ggrepel" package (version:0.9.1) [54].

Back and front-end design

The Microproteins mining database provides a user-friendly web interface that enables users to search and retrieve microprotein-stress function associations in the database (Fig. 1. and Fig. 2.). All data in the Microproteins mining database were stored and managed using MySQL (version 5.5). The web interfaces and services were built using Tomcat 8, JDK 1.8, and Bootstrap 3. Some exemplary use cases showing the utility of Mip-mining are available at https://weilab.sjtu.edu.cn/mipmining/help.

Utility and discussion

Architecture of Mip-mining

The schematic overview of the data acquisition and construction of the Mip-mining database is shown in Fig. 3. Firstly, data are collected from the GEO database with keyword searching; after a standardized RNA-seq analysis using Hisat2-stringTie-ballgown processing on HPC (High-Performance Computing) [55], the differential expression matrix is obtained. Then R packages are used for searching differentially expressed genes enrichment

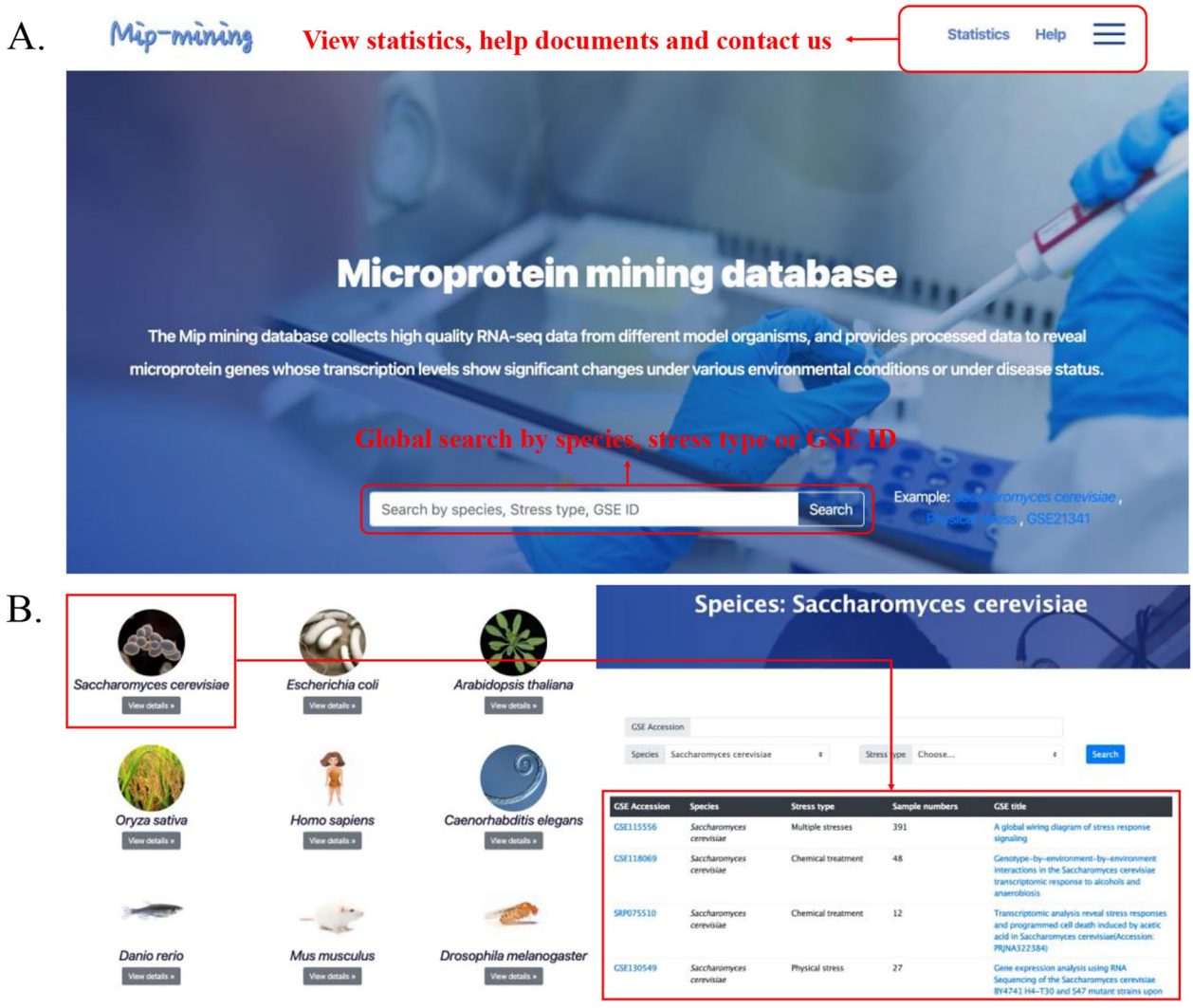

Fig. 1 User interface of the Mip-mining database. (**A**) Global search function and related information are provided on the home page. Mip-mining offers a platform to search by species, stress type, and GSE ID. (**B**) Browse specific species data from Mip-mining

analysis and result visualization. All results can be downloaded locally.

The Mip-mining database benefits the establishment of the relationship between differential expression of microproteins and various conditions (including external environmental response and internal disease development). It would help to mine the corresponding functions of microproteins. Mip-mining provide three major functions: (i) Browse and search primary data through condition type, species, and GSE accession (Fig. 1. and Fig. 2.); (ii) Identification of differentially expressed microproteins and corresponding functional enrichment analysis (Fig. 4. and Fig. 5.); (iii) Result visualization, and download.

We next demonstrate the utility of Mip-mining in studies of microprotein functions in several species through the case studies below.

Case studies

Case study 1. Stress tolerance-related microproteins in budding yeast S. cerevisiae

Yeast is commonly used in industries for food production, pharmaceutical research, chemical fermentation, and renewable energy production [56]. During bioproduction, yeast cells are subject to various stress conditions. For example, biorefinery of lignocellulosic biomass using yeast is negatively affected by decreased growth and metabolism due to inhibitors in biomass hydrolysate [57]. Among the inhibitors, acetic acid is commonly present and is highly toxic to yeast cells [58]. Improvement of acetic acid tolerance is thus desirable to develop yeast strains for efficient lignocellulosic biorefinery. In this regard, we used Mip-mining to analyze the expression of small proteins under acetic acid stress. We found that among the small proteins ranked in the GSE52160 data

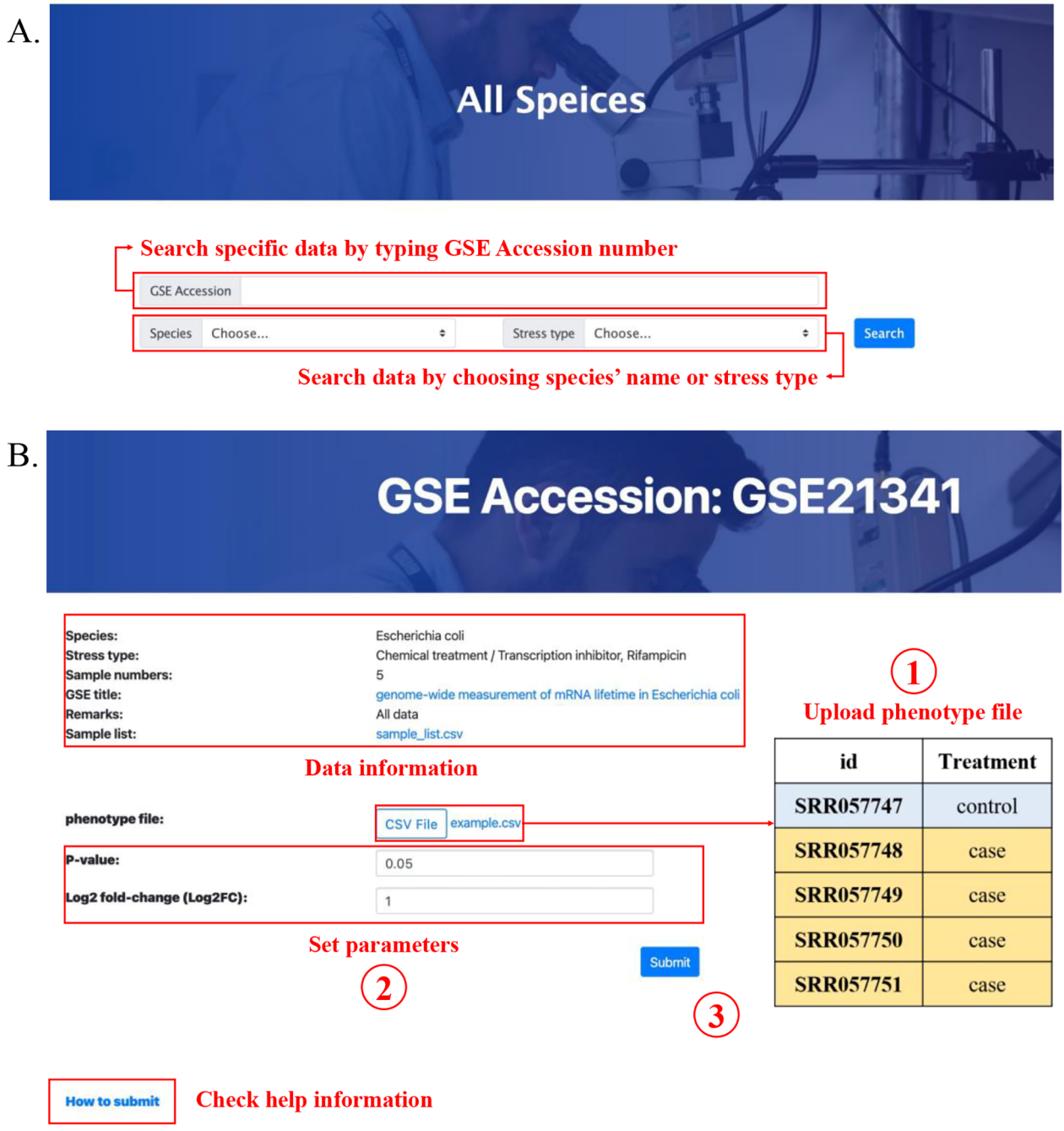

Fig. 2 User interface of the Mip-mining database. (**A**) Select search methods on the all-species page. (**B**) Personalized analysis on the analysis page. This analysis needs uploading the corresponding phenotype file, setting analyzed parameters, and then clicking the submit button. Users can also click "How to submit" to check help information

set analysis, three genes encoding microproteins showed significant changes (Table 2). As recorded in the *Saccharomyces* Genome Database (SGD) [59], the knockout of the microprotein gene *PMP2* directly affected growth under acetic acid stress, which supports that our database is functional in revealing microproteins with known roles. Furthermore, deleting *ATP15* and *SDH6* were reported to affect the growth of *S. cerevisiae* under low pH conditions and respiratory growth [60–62], respectively. The changed transcription level by the Mip-mining analysis indicates that *ATP*15 may be involved in acetic acid stress. To further examine whether *ATP15* is involved in acetic acid stress response, we overexpressed this gene using a high copy number plasmid pJFE3.

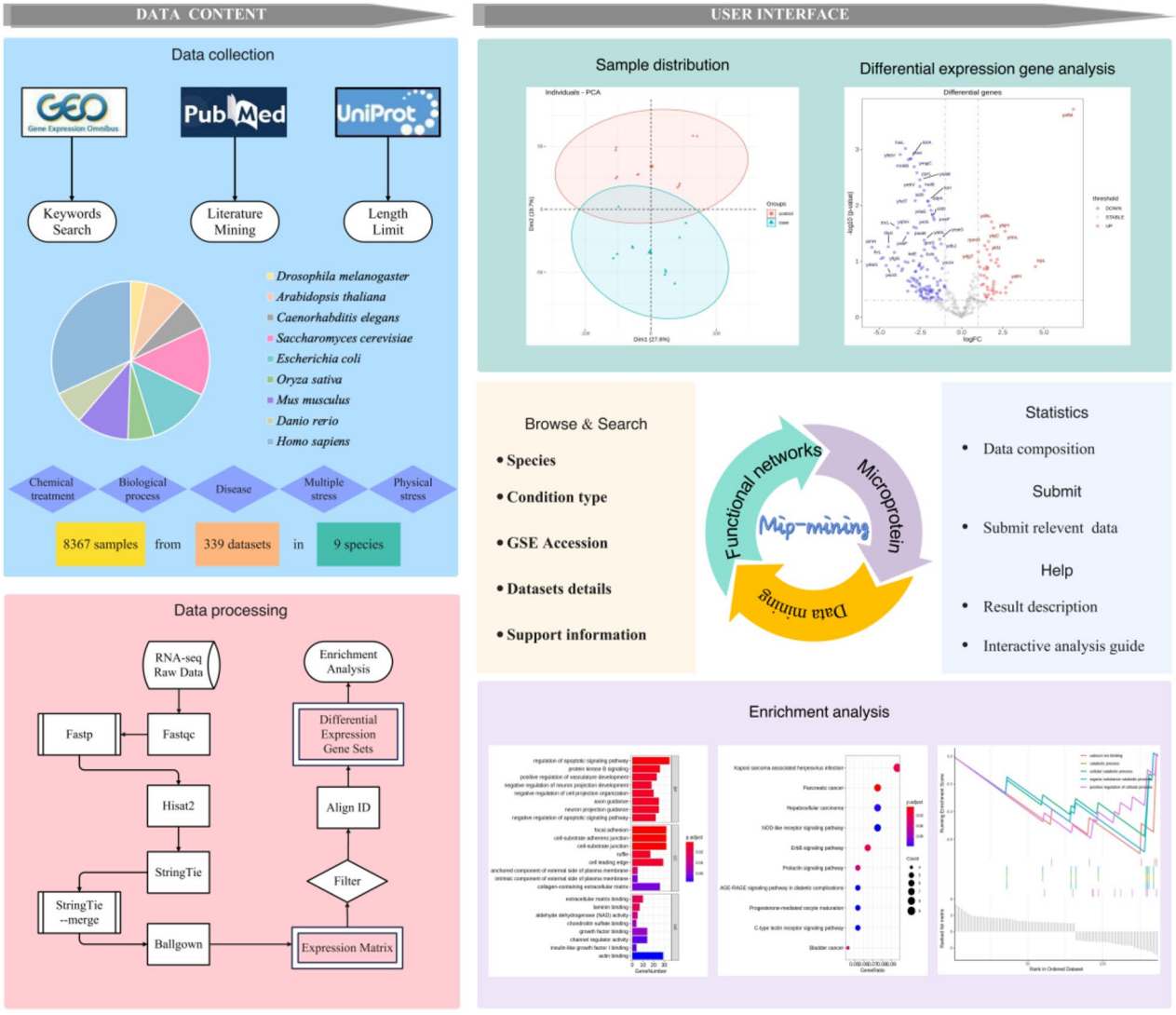

Fig. 3 Schematic overview of the data acquisition and construction of the Mip-mining database. Items with blue and background represent data sources and data processing, respectively, whereas those with a light green background indicate sample distribution and differential expression gene analysis results. In addition, items with an off-white background correspond to the browse and search methods. The content of the statistics, submit, and help pages are shown in light blue, while the purple section represents enrichment analysis results

The construction of ATP15 expressional plasmids and strains

The *ATP15* expression plasmid pJFE3 was constructed by introducing the *ATP15* gene cloned from *Saccharomyces cerevisiae* S288C into the high copy plasmid pJFE3 [63] between the sites of *TEF1p* and *PGK1t*. Then, the expression plasmid and the empty plasmid pJFE3 were transformed into the auxotroph *S. cerevisiae* BY4741 to produce the *ATP15* expressional strain BY4741-pJFE3-ATP15. The correctness of the expression plasmid and recombinant strain construction was verified by sequencing after PCR amplification. The primers used in PCR for construction and verification are listed in Table S1.

Allocation of medium

SC-Ura fluid nutrient medium: YNB(Yeast nitrogen base without amino acid with $(NH_4)_2SO_4$)6.7 g/L, amino acid mixture without URA 0.77 g/L, and glucose 20 g/L. The prepared medium was sterilized by autoclave at 115℃ for 15 min. Acid stress medium is based on SC-Ura fluid nutrient medium, adding 1 M HCL to adjust the pH to 2.3. Acetic acid stress medium is prepared based on SC-Ura fluid nutrient medium, with the addition of 4.2 g/L acetic acid.

Strain inoculation and culture

The constructed overexpressed strain and control strain were reactivated by SC-Ura liquid medium two times,

P value = 0.05, Log2FC value = 1, Job id = GSE21341_1661755050881

Download the full result: Microprotein_results.csv

Table 1. Microprotein_results.csv

ID	Log2FC	AveExpr	T.statistic	P.value	Adj.P.val	Log.odds	Group
yefM	6.7872	6.7872	10.3353	2e-04	0.4138	0.2838	UP
hisL	-3.4066	-3.4066	-7.2257	0.001	0.4138	-0.5114	DOWN
ybbV	-3.725	-3.725	-6.8369	0.0012	0.4138	-0.6583	DOWN
ldrA	-3.0968	-3.0968	-6.5686	0.0015	0.4138	-0.7682	DOWN
yoal	-3.0942	-3.0942	-6.5631	0.0015	0.4138	-0.7706	DOWN
mokB	-3.2358	-3.2358	-6.5092	0.0015	0.4138	-0.7936	DOWN
ymgC	-2.8717	-2.8717	-6.0912	0.002	0.441	-0.983	DOWN
pyrL	-2.7084	-2.7084	-5.7447	0.0026	0.5408	-1.1559	DOWN
ysaB	-2.5292	-2.5292	-5.3646	0.0035	0.6241	-1.3639	DOWN

Fig. 4 The first part of the Mip-mining database results page. DGEs (Differential Gene Expression) result table contains the microprotein gene names, Log2FC: an estimate of the log2-fold-change corresponding to the contrast (case vs. control), AveExpr: average log2-expression for the sample, T.statistic: moderated t-statistic, P.value: raw p-value, Adj.P.val: adjust corrected p-value, Log.odds: log-odds that the gene is differentially expressed and Group: gene label indicates up-regulation or down-regulation or stabilization of microprotein visualization of the sample distribution

which were then added to SC-Ura fluid nutrient medium to obtain seed liquid. The activated strains were inoculated into shake bottles with the initial OD600 of 0.03, cultured at 30 °C shaking at 150 rpm. The broth was sampled at an appropriate time point to detect the growth under stress-free and stress conditions.

The results revealed that high-level expression of *ATP15* severely inhibits growth in the presence of acetic acid; about 24 h longer lag phase time was observed when *ATP15* was overexpressed. Reduced biomass was observed under non-stress and low pH (2.3) conditions. The unprecedented growth repression by *ATP15* overexpression under acetic acid stress confirmed that this protein is critical in combating stress (Fig. 6.).

Case study 2. Microproteins in the model plant A. thaliana

Plant stress responses have been studied to provide a basis for breeding crops that resist salt, cold environment,

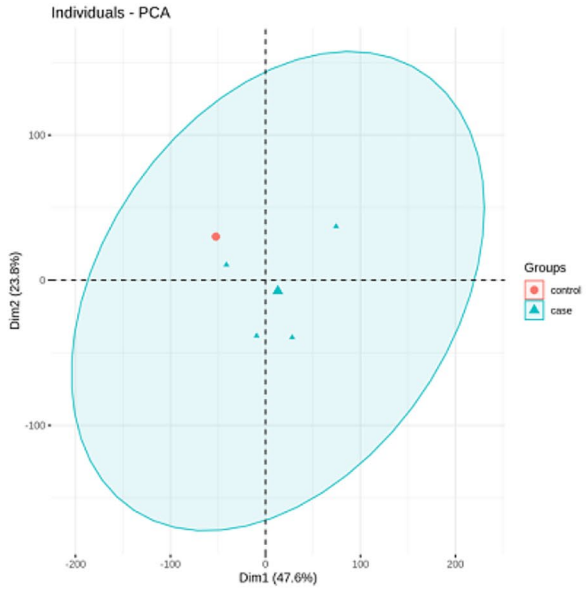

Figure1. all_samples_PCA

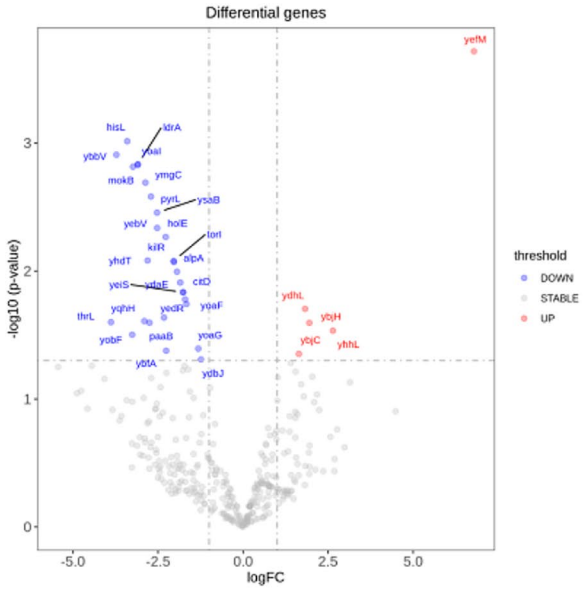

Figure2. VolcanoMP

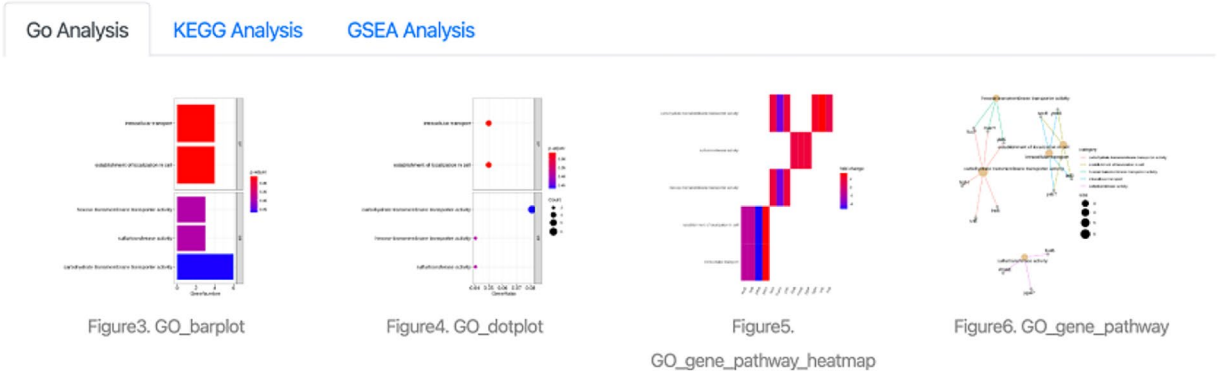

Figure3. GO_barplot

Figure4. GO_dotplot

Figure5. GO_gene_pathway_heatmap

Figure6. GO_gene_pathway

Download the other results: GO_results.csv, KEGG_results.csv

Fig. 5 The second part of the result page. The page contains a visualization of sample distribution and DGEs (Differential Gene Expression) results, as well as the enrichment analysis results integrating GO, KEGG, and GSEA parts from RNA-seq data, and it also provides a download link for GO and KEGG analysis outcomes

drought, or microbial pathogens [18]. The temperature is an essential factor among these stress conditions encountered by plants. Low or high temperatures affect the development of plants and their immunity to harsh conditions [64]. In this regard, we select GSE116004 for analysis, which compares the global transcription of the model plant *A. thaliana* at 37 °C with the control condition at normal temperature (Table S1). We observed changes in *PIP1* and *PIP2*, which were annotated as endogenous secreted peptides that elicit an immune response and positive regulators of defence response [65]. So far, no reports have been found on the functions of these two proteins in heat resistance. Therefore, our results revealed the plant microproteins' potential that can be further investigated for their functions under specific environmental conditions.

Case study 3. Microproteins related to human cancer

Breast cancer is a severe threat to women's health, and triple-negative breast cancer is challenging to treat due

Table 2 Yeast microproteins identified by Mip-mining by analyzing the dataset of GSE52160*

Gene name	Log_2FC	Phenotype and function
PMP2	1.94	Acetic acid resistance decreased by gene deletion
ATP15	-2.92	Propionic acid pH resistance decreased by gene deletion
SDH6	-3.4	Respiratory growth is absent after adding 2% acetate by deletion

*Functions were retrieved from *Saccharomyces* Genome Database (SGD), https://www.yeastgenome.org/

to its lack of therapeutic targets, high recurrence rate, and uncomplicated metastasis. We selected the dataset GSE171957 to study the connection between microproteins and triple-negative breast cancer, hoping to provide more therapeutic directions for triple-negative breast cancer from the perspective of microproteins (Table S2). According to the results of the Mip-mining analysis, we conducted a literature survey and found that PKIB is involved in the signaling pathway induced by cAMP [66]. CENPW is associated with nucleosomes [67]. COA4 [68] is associated with cytochrome c oxidase. Among significantly down-regulated genes, long non-coding RNA SNHG12 has been proven to be a potential pan-cancer marker and therapeutic target [69]. NUPR1 promotes cancer cell metastasis, can help cancer cells adapt to the microenvironment after chemotherapy and play a role in drug resistance [70]. In addition, reducing RPS27L can regulate autophagy and promote tumorigenesis [71].In addition to microproteins directly associated with triple-negative breast cancer, we also found that significant downregulation of DPY30, which is thought to regulate the epithelial-mesenchymal transition to affect cervical squamous cell carcinoma [72], and is so far an unexplored microprotein regulator.

To summarize, through case studies of triple-negative breast cancer, we can find relevant key regulators that have been proven and can also provide researchers with more potential therapeutic targets and research directions.

Discussion

Mip-mining in the current study is the first database focusing on transcriptome profiles in microproteins related to environmental stress tolerance or diseases. It will be useful for researching and applying microproteins in sustainable bioproduction, biomarker discovery, and disease treatment. Compared with the existing microprotein databases, including SmProt, sORFs.org, and PsORF, among others [23–31] contributing to the widespread existence of microproteins in living organisms, Mip-mining is unique because it aims to reveal the effects of microproteins under a wide range of conditions. The database contains expanded data set from more diverse organisms, which includes microorganisms, plants, and animals. Additionally, the data we collected focus on multiple stress conditions and various diseases, which enables the exploration of microproteins with essential functions. Besides, only high-quality transcriptomic data were collected, and most of the RNA-seq data have literature support for easy traceability, which guarantees the

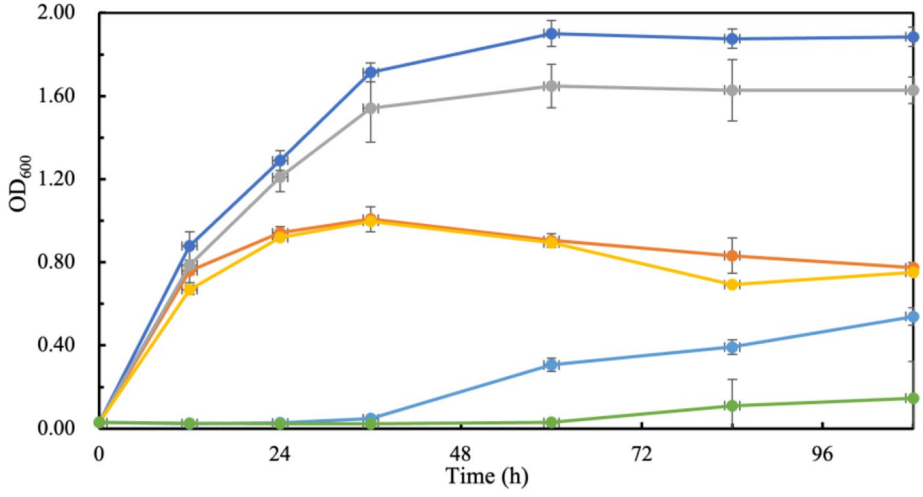

Fig. 6 Overexpression of *ATP15* affects yeast growth and stress tolerance. Dark blue and dark orange, growth of the control strain and the *ATP15* overexpression strain carrying the empty plasmid and the *ATP15* expression plasmid, respectively, under stress-free conditions; Grey and light orange, growth of the control strain and the *ATP15* overexpression strain at pH 2.3; Light blue and green, growth of the control strain and the *ATP15* overexpression strain in the presence of 4.2 g/L acetic acid. Yeast strains were grown in YPD broth at 30 shaking at 150 rpm with or without addition of acetic acid, and pH 2.3 was adjusted using 1 M HCl.

reliability of the analysis. Although most other databases collect data based on mass spectrometry analysis and ribosome profiling for microprotein studies, we emphasize that the transcription of microprotein genes contains essential information and cannot be neglected. Firstly, transcription regulation starts gene expression, and the co-transcription of microproteins and other genes correlates with their functions. Secondly, so far, detection of the translation of microproteins is still restricted by technical limitations due to low expression and or specific properties of microproteins; therefore, transcriptome data are a critical complement for in-depth studies.

The Mip-mining database establishes the connection between environmental stress or disease, microproteins, and functional characterization. Through analysis, it is possible to quickly clarify the changes in the mRNA level in the specific organism under each stress/disease condition, supported by multiple data sets. Enrichment analysis can help users to deduce which pathways are more important under certain conditions, and the data can be used to trace back which pathways small proteins are involved in. Compared with other related databases, our current database is more beneficial for researchers to establish functional exploration and design experiments for further mechanism studies.

The role of microproteins as regulatory proteins in various living organisms is increasingly recognized [73, 74]. However, studies on microproteins should not ignore the synergistic effects of these essential proteins with other proteins, such as the differential expression of multiple proteins simultaneously. Mip-mining provides a novel platform to explore protein interaction networks under various stressful environments involving microproteins. The information provided by our database can be further used to study protein interaction networks to design more powerful small proteins. In this regard, the results may help employ microproteins to assist large protein complexes in various life activities.

We provide the function of screening differentially expressed microproteins for each set of data, but the information supplement for each microprotein has not yet been completed. Links with other reference microprotein databases can supplement more microprotein-related information. Up to now, Mip-mining contains information about microproteins related to stress conditions in 9 species. With the emergence of more RNA-seq data from non-model organisms and the improvement and advancement of sequencing technology, we will continue to collect microprotein information of more other species and refine related external conditions, for example, more data related to various other human diseases.

Conclusion

We present the Mip-mining database - an innovative tool that allows users to conduct personalized analysis of microprotein functions. The Mip-mining database hosts 336 sets of high-quality transcriptome data from 8626 samples and nine representative living organisms, including microorganisms, plants, animals, and humans. Microproteins are potentially related to various diseases and environmental stress conditions, including chemical, physical, biological, and multiple stresses, and thus understanding a related microprotein or set of microproteins is crucial for a thorough understanding of these conditions. Users can select specific cutoff values for enhanced customization of their analysis. Consequently, this tool serves as a valuable resource for research communities investigating microproteins in diverse scientific fields.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09735-1.

Supplementary Material 1

Acknowledgements

The authors are grateful to the graduate student Mingming Jiang and the undergraduate students Junchen Yao and Yuxiang Zhang at Shanghai Jiao Tong University that contribute to collecting the datasets for this work.

Authors' contributions

X.Z., D.W., and Y.W. designed the work and organized the research team, reviewed and revised the manuscript, B.Z., J.Z., M.W., W.W. completed data collection. B.Z. and J.Z. completed data preprocessing and backend construction. YG completed the visualization of Mip-mining. M.W. completed the experimental verification. B.Z., J.Z., A.M., D.W., X.Z. and Y.W. wrote the main manuscript text and figures. Y.X. and S.L. optimized the running speed of the database. All authors read and approved the final manuscript.

Funding

This work was supported by the State Key Research and Development Program (No. 2022YFE0108500) and grant from the State Key Laboratory of Microbial Metabolism (Shanghai Jiao Tong Unversity). Dong-Qing Wei is supported by grants from the National Science Foundation of China (Grant No. 32070662, 61832019, 32030063), Intergovernmental International Scientific and Technological Innovation and Cooperation Program of The National Key R&D Program (2023YFE0199200). Y. Wang thanks support from the grants from the National Natural Science Foundation of China (No.32200531) and Startup Fund for Young Faculty at SJTU (SFYF at SJTU). The computations were partially performed at the Pengcheng Lab and the Center for High-Performance Computing, Shanghai Jiao Tong University.

Data Availability

All data, including preprocessed transcriptome data and filtered microprotein information are stored on the Mip-mining website (https://weilab.sjtu.edu.cn/mipmining/).

Code Availability

Code for Mip-mining filtering is available at https://github.com/GlancerZ/Mipmining.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 11 February 2023 / Accepted: 12 October 2023
Published online: 02 November 2023

References

1. Couso JP, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. 2017;18(9):575–89.
2. Khitun A, Ness TJ, Slavoff SA. Small open reading frames and cellular stress responses. Mol Omics. 2019;15(2):108–16.
3. Schlesinger D, Elsässer SJ. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. Febs j. 2022;289(1):53–74.
4. Orr MW, Mao Y, Storz G, Qian SB. Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res. 2020;48(3):1029–42.
5. Ji X, Cui C, Cui Q. smORFunction: a tool for predicting functions of small open reading frames and microproteins. BMC Bioinformatics. 2020;21(1):455.
6. Durrant MG, Bhatt AS. Automated prediction and annotation of Small Open Reading frames in Microbial genomes. Cell Host Microbe. 2021;29(1):121–131e124.
7. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A. Accurate annotation of human protein-coding small open reading frames. Nat Chem Biol. 2020;16(4):458–68.
8. Mat-Sharani S, Firdaus-Raih M. Computational discovery and annotation of conserved small open reading frames in fungal genomes. BMC Bioinformatics. 2019;19(Suppl 13):551.
9. Cao X, Khitun A, Luo Y, Na Z, Phoodokmai T, Sappakhaw K, Olatunji E, Uttamapinant C, Slavoff SA. Alt-RPL36 downregulates the PI3K-AKT-mTOR signaling pathway by interacting with TMEM24. Nat Commun. 2021;12(1):508.
10. Wu Q, Kuang K, Lyu M, Zhao Y, Li Y, Li J, Pan Y, Shi H, Zhong S. Allosteric deactivation of PIFs and EIN3 by microproteins in light control of plant development. Proc Natl Acad Sci U S A. 2020;117(31):18858–68.
11. Guo X, Chavez A, Tung A, Chan Y, Kaas C, Yin Y, Cecchi R, Garnier SL, Kelsic ED, Schubert M, et al. High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast. Nat Biotechnol. 2018;36(6):540–6.
12. Impens F, Rolhion N, Radoshevich L, Bécavin C, Duval M, Mellin J, García Del Portillo F, Pucciarelli MG, Williams AH, Cossart P. N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in Listeria monocytogenes. Nat Microbiol. 2017;2:17005.
13. Kang M, Tang B, Li J, Zhou Z, Liu K, Wang R, Jiang Z, Bi F, Patrick D, Kim D, et al. Identification of miPEP133 as a novel tumor-suppressor microprotein encoded by miR-34a pri-miRNA. Mol Cancer. 2020;19(1):143.
14. Wang G, Zietz CM, Mudgapalli A, Wang S, Wang Z. The evolution of the antimicrobial peptide database over 18 years: milestones and new features. Protein Sci. 2022;31(1):92–106.
15. Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, Cavalheiro M, Antunes M, Lemos A, Pedreira T, et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucleic Acids Res. 2018;46(D1):D348–d353.
16. dos Santos SC, Sá-Correia I. Yeast toxicogenomics: lessons from a eukaryotic cell model and cell factory. Curr Opin Biotechnol. 2015;33:183–91.
17. Thorwall S, Schwartz C, Chartron JW, Wheeldon I. Stress-tolerant non-conventional microbes enable next-generation chemical biosynthesis. Nat Chem Biol. 2020;16(2):113–21.
18. Rivero RM, Mittler R, Blumwald E, Zandalinas SI. Developing climate-resilient crops: improving plant tolerance to stress combination. Plant J. 2022;109(2):373–89.
19. Ghosh A, Shcherbik N. Effects of oxidative stress on protein translation: implications for Cardiovascular Diseases. Int J Mol Sci 2020, 21(8).
20. Gaillard H, García-Muse T, Aguilera A. Replication stress and cancer. Nat Rev Cancer. 2015;15(5):276–89.
21. Lam FH, Ghaderi A, Fink GR, Stephanopoulos G. Biofuels. Engineering alcohol tolerance in yeast. Science. 2014;346(6205):71–5.
22. Bhati KK, Blaakmeer A, Paredes EB, Dolde U, Eguen T, Hong SY, Rodrigues V, Straub D, Sun B, Wenkel S. Approaches to identify and characterize microProteins and their potential uses in biotechnology. Cell Mol Life Sci. 2018;75(14):2529–36.
23. Hazarika RR, De Coninck B, Yamamoto LR, Martin LR, Cammue BP, van Noort V. ARA-PEPs: a repository of putative sORF-encoded peptides in Arabidopsis thaliana. BMC Bioinformatics. 2017;18(1):37.
24. Chen Y, Li D, Fan W, Zheng X, Zhou Y, Ye H, Liang X, Du W, Zhou Y, Wang K. PsORF: a database of small ORFs in plants. Plant Biotechnol J. 2020;18(11):2158–60.
25. Hao Y, Zhang L, Niu Y, Cai T, Luo J, He S, Zhang B, Zhang D, Qin Y, Yang F, et al. SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. Brief Bioinform. 2018;19(4):636–43.
26. Li Y, Zhou H, Chen X, Zheng Y, Kang Q, Hao D, Zhang L, Song T, Luo H, Hao Y et al. SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genomics, Proteomics & Bioinformatics* 2021.
27. Brunet MA, Brunelle M, Lucier JF, Delcourt V, Levesque M, Grenier F, Samandi S, Leblanc S, Aguilar JD, Dufour P, et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res. 2019;47(D1):D403–d410.
28. Brunet MA, Lucier JF, Levesque M, Leblanc S, Jacques JF, Al-Saedi HRH, Guilloy N, Grenier F, Avino M, Fournier I, et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res. 2021;49(D1):D380–d388.
29. Wan J, Qian SB. TISdb: a database for alternative translation initiation in mammalian cells. Nucleic Acids Res. 2014;42(Database issue):D845–850.
30. Olexiouk V, Crappé J, Verbruggen S, Verhegen K, Martens L, Menschaert G. sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2016;44(D1):D324–329.
31. Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2018;46(D1):D497–d502.
32. Heo HS, Lee S, Kim JM, Choi YJ, Chung HY, Oh SJ. tsORFdb: theoretical small open reading frames (ORFs) database and massProphet: peptide mass fingerprinting (PMF) tool for unknown small functional ORFs. Biochem Biophys Res Commun. 2010;397(1):120–6.
33. F RC, Vasconcelos ATR. OCCAM: prediction of small ORFs in bacterial genomes by means of a target-decoy database approach and machine learning techniques. *Database (Oxford)* 2020, 2020.
34. Guruceaga E, Garin-Muga A, Segura V. MiTPeptideDB: a proteogenomic resource for the discovery of novel peptides. Bioinformatics. 2020;36(1):205–11.
35. Dhamija S, Menon MB. Non-coding transcript variants of protein-coding genes - what are they good for? RNA Biol. 2018;15(8):1025–31.
36. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: where chemistry meets biology. Science. 2020;367(6476):392–6.
37. Montaño López J, Duran L, Avalos JL. Physiological limitations and opportunities in microbial metabolic engineering. Nat Rev Microbiol. 2022;20(1):35–48.
38. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41(Database issue):D991–995.
39. Pitocco D, Zaccardi F, Di Stasio E, Romitelli F, Santini SA, Zuppi C, Ghirlanda G. Oxidative stress, nitric oxide, and Diabetes. Rev Diabet Stud. 2010;7(1):15–25.
40. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2018;47(D1):D766–73.
41. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754–d761.
42. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–.

43. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19–21.
44. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
45. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
46. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
47. Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. Ballgown bridges the gap between transcriptome assembly and expression analysis. Nat Biotechnol. 2015;33(3):243–6.
48. Mundt AKaF. : (2020).factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRANR-project.org/package=factoextra
49. Sebastien Le JJ. FactoMineR: an R Package for Multivariate Analysis. Journal of Statistical Software. J Stat Softw. 2008;25(1):1–18.
50. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
51. Yu G. (2019). enrichplot: Visualization of Functional Enrichment Result. R package version 1.6.1. https://github.com/GuangchuangYu/enrichplot.
52. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics. 2012;16(5):284–7.
53. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016.
54. Slowikowski K. (2021). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel.
55. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164–5.
56. Guirimand G, Kulagina N, Papon N, Hasunuma T, Courdavault V. Innovative tools and strategies for optimizing yeast cell factories. Trends Biotechnol. 2021;39(5):488–504.
57. Zhang MM, Chen HQ, Ye PL, Wattanachaisaereekul S, Bai FW, Zhao XQ. Development of robust yeast strains for lignocellulosic biorefineries based on genome-wide studies. Prog Mol Subcell Biol. 2019;58:61–83.
58. Guaragnella N, Bettiga M. Acetic acid stress in budding yeast: from molecular mechanisms to applications. Yeast. 2021;38(7):391–400.
59. Lang OW, Nash RS, Hellerstedt ST, Engel SR. An introduction to the Saccharomyces Genome Database (SGD). Methods Mol Biol. 2018;1757:21–30.
60. Na U, Yu W, Cox J, Bricker DK, Brockmann K, Rutter J, Thummel CS, Winge DR. The LYR factors SDHAF1 and SDHAF3 mediate maturation of the iron-sulfur subunit of succinate dehydrogenase. Cell Metab. 2014;20(2):253–66.
61. Kawahata M, Masaki K, Fujii T, Iefuji H. Yeast genes involved in response to lactic acid and acetic acid: acidic conditions caused by the organic acids in Saccharomyces cerevisiae cultures induce expression of intracellular metal metabolism genes regulated by Aft1p. FEMS Yeast Res. 2006;6(6):924–36.
62. Mira NP, Lourenço AB, Fernandes AR, Becker JD, Sá-Correia I. The RIM101 pathway has a role in Saccharomyces cerevisiae adaptive response and resistance to propionic acid and other weak acids. FEMS Yeast Res. 2009;9(2):202–16.
63. Shen Y, Chen X, Peng B, Chen L, Hou J, Bao X. An efficient xylose-fermenting recombinant Saccharomyces cerevisiae strain obtained through adaptive evolution and its global transcription profile. Appl Microbiol Biotechnol. 2012;96(4):1079–91.
64. Ding Y, Shi Y, Yang S. Molecular regulation of plant responses to environmental temperatures. Mol Plant. 2020;13(4):544–64.
65. Rahman A, Kawamura Y, Maeshima M, Rahman A, Uemura M. Plasma membrane aquaporin members PIPs Act in Concert to regulate cold acclimation and freezing tolerance responses in Arabidopsis thaliana. Plant Cell Physiol. 2020;61(4):787–802.
66. Zhang JB, Song W, Wang YY, Liu MG, Sun MM, Liu H. Study on correlation between PKIB and pAkt expression in Breast cancer tissues. Eur Rev Med Pharmacol Sci. 2017;21(6):1264–9.
67. Wang L, Wang H, Yang C, Wu Y, Lei G, Yu Y, Gao Y, Du J, Tong X, Zhou F et al. Investigating CENPW as a Novel Biomarker Correlated with the development and poor prognosis of breast carcinoma. Front Genet 2022, 13.
68. Kwon YS, Lee MG, Baek J, Kim NY, Jang H, Kim S. Acyl-CoA synthetase-4 mediates radioresistance of Breast cancer cells by regulating FOXM1. Biochem Pharmacol. 2021;192:114718.
69. Tamang S, Acharya V, Roy D, Sharma R, Aryaa A, Sharma U, Khandelwal A, Prakash H, Vasquez KM, Jain A. SNHG12: an LncRNA as a potential therapeutic target and biomarker for Human Cancer. Front Oncol. 2019;9:901.
70. Wang L, Sun J, Yin Y, Sun Y, Ma J, Zhou R, Chang X, Li D, Yao Z, Tian S, et al. Transcriptional coregualtor NUPR1 maintains tamoxifen resistance in Breast cancer cells. Cell Death Dis. 2021;12(2):149.
71. Xiong X, Liu X, Li H, He H, Sun Y, Zhao Y. Ribosomal protein S27-like regulates autophagy via the β-TrCP-DEPTOR-mTORC1 axis. Cell Death Dis. 2018;9(11):1131.
72. Li J, Zhou P, Xiong C, Hoi SC. Prototypical contrastive learning of unsupervised representations. arXiv Preprint arXiv:200504966 2020.
73. Bhati KK, Dolde U, Wenkel S. MicroProteins: expanding functions and novel modes of regulation. Mol Plant. 2021;14(5):705–7.
74. Wu QQ, Zhong SW, Shi H. MicroProteins: Dynamic and accurate regulation of protein activity.

Publisher's Note