Plant Physiology®

Plant Physiology, 2024, 00, 1–10
https://doi.org/10.1093/plphys/kiae311
Advance access publication 29 May 2024
Research Article

NCPbook: A comprehensive database of noncanonical peptides

Abdul Sami,[1,†] Mengjia Fu,[2,†] Haoqiang Yin,[1,†] Usman Ali,[1] Lei Tian,[1] Shunxi Wang,[1] Jinghua Zhang,[1] Xueyan Chen,[1] Hehuan Li,[1] Minghui Chen,[1] Wen Yao,[2,*] and Liuji Wu[1,*]

[1]National Key Laboratory of Wheat and Maize Crop Science, College of Agronomy, Henan Agricultural University, Zhengzhou 450046, China
[2]National Key Laboratory of Wheat and Maize Crop Science, College of Life Sciences, Henan Agricultural University, Zhengzhou 450046, China
*Author for correspondence: wlj200120@163.com (L.W.), yaowen@henau.edu.cn (W.Y.)
†These authors contributed equally.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/general-instructions) is: Liuji Wu (wlj200120@163.com) and Wen Yao (yaowen@henau.edu.cn).

Abstract

Noncanonical peptides (NCPs) are a class of peptides generated from regions previously thought of as noncoding, such as introns, 5′ UTRs, 3′ UTRs, and intergenic regions. In recent years, the significance and diverse functions of NCPs have come to light, yet a systematic and comprehensive NCP database remains absent. Here, we developed NCPbook (https://ncp.wiki/ncpbook/), a database of evidence-supported NCPs, which aims to provide a resource for efficient exploration, analysis, and manipulation of NCPs. NCPbook incorporates data from diverse public databases and scientific literature. The current version of NCPbook includes 180,676 NCPs across 29 different species, evidenced by MS, ribosome profiling, or molecular experiments. These NCPs are distributed across kingdoms, comprising 123,408 from 14 plant species, 56,999 from 7 animal species, and 269 from 8 microbial species. Furthermore, NCPbook encompasses 9,166 functionally characterized NCPs playing important roles in immunity, stress resistance, growth, and development. Equipped with a user-friendly interface, NCPbook allows users to search, browse, visualize, and retrieve data, making it an indispensable platform for researching NCPs in various plant, animal, and microbial species.

Introduction

Peptides are small biological compounds comprised of 2 to 100 amino acid (AA) residues that play critical roles in various biological processes (Tavormina et al. 2015). Noncanonical peptides (NCPs) are a class of peptides derived from noncanonical ORF regions such as introns, 5′ UTRs, 3′ UTRs, and intergenic regions. As a distinct class of peptides, NCPs have gained substantial attention due to their increasing recognition as biologically significant entities in various organisms. For instance, microRNA-encoded peptide858a (miPEP858a), microRNA-encoded peptide165a (miPEP165a), microRNA-encoded peptide 171b (miPEP171b), *Vitis vinifera* microRNA-encoded peptide171d (vvi-miPEP171d), and microRNA-encoded peptide172c (miPEP 172c) in plants (Lauressergues et al. 2015; Couzigou et al. 2016; Sharma et al. 2020; Chen, Deng, et al. 2020) and AW112010, a peptide located in endoplasmic reticulum, and myoregulin in human and animals (Anderson et al. 2015; Jackson et al. 2018; Sun et al. 2021) have been reported to play significant roles in various biological processes, including growth, development, translational regulation, and environmental responses. In our previous study, we developed an integrated peptidogenomic pipeline for the large-scale discovery of plant NCPs and identified antifungal activity conferred by NCPs (Wang et al. 2020; Tian et al. 2021).

Moreover, a plethora of evidence from a diverse range of studies in plants and animals has revealed that NCPs can indeed be encoded by noncoding RNAs (ncRNAs), including long ncRNAs (lncRNAs), microRNAs (miRNAs), circular RNAs (circRNAs), and small nuclear RNAs (Wang, Mao, and Liu 2019; Ormancey et al. 2021; Vale et al. 2021; Xing et al. 2021; Lauressergues et al. 2022; Luo et al. 2022). It is worth noting that an increasing number of NCPs derived from ncRNAs have been investigated for their significant roles in various physiological and biological processes. For example, an NCP (miPEP165a) encoded by pri-miR165a plays a crucial role in root development in *Arabidopsis* (*Arabidopsis thaliana*) (Lauressergues et al. 2015). In *Drosophila* (*Drosophila melanogaster*), 4 NCPs encoded by an lncRNA are required for epidermal differentiation (Kondo et al. 2010).

Biological databases are comprehensive libraries of biological molecules collected from high-throughput experiments, computational analysis, and published articles. They play a crucial role in assisting scientists in analyzing and explaining various biological phenomena, including growth, development, stress response, and metabolism (Baxevanis and Bateman 2015; Boschiero et al. 2020). With the increasing interest in NCPs, several novel NCPs have been discovered and characterized. Nevertheless, researchers still face significant challenges when probing the functional roles of NCPs due to the dispersion of relevant information throughout the existing literature. Previously, several databases like small protein (SmProt), small peptides encoded by noncoding RNAs in cancer patients (SPENCER), and plant ORF (PsORF) have documented peptides derived from ncRNAs or small ORFs (sORFs) (Chen, Li, et al. 2020; Li et al. 2021; Luo et al. 2022). While these databases provide valuable information, most of them contain limited numbers of NCPs derived from a few types of noncoding regions in a small number of species. Additionally, some of these databases archived predicted NCPs without supporting evidence. Therefore, there remains

Received January 30, 2024. Accepted May 10, 2024.

a need for a comprehensive database serving as centralized repository of NCPs for modern biological research.

In this perspective, we developed a comprehensive database, NCPbook (https://ncp.wiki/ncpbook/), to facilitate the exploration of NCPs from a variety of noncoding regions, including introns, intergenic regions, upstream ORFs (uORFs), downstream ORFs (dORFs), lncRNAs, miRNAs, circRNAs, and ribosomal RNAs (rRNAs) across 29 distinct species of plants, animals, and microbes. The current version of NCPbook encompasses a total of 180,676 NCPs, primarily evidenced by MS, ribosome profiling (Ribo-seq), or molecular experiments (MEs) such as western blot, in vitro translation assays, and ELISA. NCPbook will be significant for the research community as it not only streamlines data access but also promotes collaboration and knowledge sharing among researchers studying NCPs. In summary, NCPbook offers a user-friendly interface for querying and browsing comprehensive information associated with NCPs and will prove instrumental in the integration, assessment, and functional exploration of NCPs in biological research.

Results

Overview of NCPbook construction

To comprehensively investigate the presence of NCPs across various species of plants, animals, and microbes, we conducted an extensive data-gathering effort. This process involved extracting NCP-related information from both scientific literature and publicly available repositories including coding and ncRNA database (cncRNAdb), functional ncRNA-encoded peptides (FuncPep), noncoding experimentally validated peptides (ncEP), putative noncoding peptide database (PncPEPDB), PsORF, SmProt, and SPENCER (Li et al. 2018, 2021; Dragomir et al. 2020; Liu et al. 2020; Chen, Li, et al. 2020; Huang et al. 2021; Luo et al. 2022). Throughout the data collection phase, we excluded NCPs that lacked substantial supporting evidence and canonical peptides documented in these databases (Fig. 1). To ensure the integration and standardized presentation of data from diverse sources within NCPbook, we applied a uniform data format to all collected datasets. Each NCP in NCPbook encompasses the following information: a unique identifier and the NCP's associated species; chromosome data (including chromosome ID, start and end coordinates, and strand); the NCP's sequence, length, and function; details regarding its host gene (such as gene name, gene ID, and gene type); database name; and relevant published literature (including title, publication date, and DOI) pertaining to the NCP.

Construction of the NCPbook database based on collection and curation of NCPs

The NCPbook database was constructed by collecting NCPs across 29 diverse species. The initial data acquisition involved the collection of 34,593 NCPs from thorough examination of over 100 research publications (Fig. 2A; Supplementary Tables S1 and S2). Among these, 6,294 NCPs were derived from plants, 28,054 NCPs from animals, and 245 NCPs from microbes (Supplementary Table S2). Notably, a proportion of these NCPs were elucidated through a robust peptidogenomic pipeline, as exemplified by the identification of 1,993 NCPs in *Zea mays* and 1,860 NCPs in *A. thaliana* reported in our prior work (Wang et al. 2020). Similarly, 1,897 NCPs from *V. vinifera* were documented in a recent study employing a peptidogenomic pipeline (Pei et al. 2022). Next, we collected NCPs from different publicly available databases including cncRNAdb, FuncPep, ncEP, PncPEPDB, PsORF, SmProt, and SPENCER (Li et al. 2018, 2021; Dragomir et al. 2020; Liu et al. 2020; Chen, Li, et al. 2020; Huang et al. 2021; Luo et al. 2022). As a result, 146,083 NCPs were retrieved from these databases, including 117,114 from plants, 28,945 from animals, and 24 from microbes (Fig. 2A; Supplementary Tables S2 and S3).

Consequently, the NCPbook database encompasses a comprehensive collection of 180,676 unique NCPs, distributed as follows: 123,408 from 14 plants, 56,999 from 7 animals, and 269 from 8 microbial species. Among the 123,408 NCPs from 14 plants, *A. thaliana* contributed the highest number (47,150 NCPs), followed by *Chlamydomonas* (*Chlamydomonas reinhardtii*) (25,884 NCPs), cotton (*Gossypium arboreum*) (20,134 NCPs), maize (*Z. mays*) (17,012 NCPs), rice (*Oryza sativa*) (10,943 NCPs), and grapes (*V. vinifera*) (2,006 NCPs). However, a limited number of NCPs were collected for the other 8 plant species in NCPbook. The plant species strawberry (*Fragaria vesca*), soybean (*Glycine max*), *Lotus japonicus*, *Medicago truncatula*, *Phaeodactylum tricornutum*, bean (*Phaseolus vulgaris*), *Physcomitrium patens*, and tomato (*Solanum lycopersicum*) contained 18, 3, 1, 11, 202, 17, 4, and 23 NCPs, respectively. It should be noted that it does not necessarily indicate a scarcity of NCPs in these 8 plant species. Rather, it is probably caused by the incomplete annotation of NCPs in these plants.

Characteristics of NCPs collected in NCPbook

In NCPbook, the collection of NCPs is categorized based on the type of noncoding regions from which they originate. Our analysis unveiled distinct counts of NCPs: 39,774 from uORFs, 48,963 from dORFs, 16,026 from intergenic regions, and 420 from introns. Apart from this, 27,576 NCPs derived from lncRNAs, 731 NCPs from circRNAs, 182 NCPs from miRNAs, and 73 NCPs from rRNAs have also been curated in NCPbook. Additionally, NCPbook also contains 46,931 NCPs originating from miscellaneous types of noncoding regions, including antisense strands and different types of junctions (Fig. 2B). All NCPs incorporated into NCPbook adhered to a criterion of being ≤100 AA in length (Fig. 2C). In fact, approximately 83% of NCPs featured in NCPbook exhibit a length shorter than 40 AA.

By classifying NCPs according to their respective species, we delineated the distribution as follows: 47,150 in *Arabidopsis*, 25,884 in *Chlamydomonas*, 20,134 in cotton, 17,012 in maize, 10,943 in rice, 31,806 in human, 24,091 in mouse, and 3,656 in other species (Fig. 2D; Supplementary Table S1). When considering the timeline of NCP discovery based on the year of their report, our analysis revealed a modest count of 999 NCPs identified between 1989 and 2013. In contrast, an impressive surge has occurred since then, with a current count of 179,677 NCPs identified to date (Fig. 2E). The NCPs in NCPbook are substantiated by various supporting evidence. We noted that 43,195 NCPs were supported by MS, 135,268 NCPs were corroborated by Ribo-seq, and 214 NCPs were validated through ME. Notably, 1,999 NCPs within NCPbook were supported by multiple lines of evidence (Fig. 2F; Supplementary Table S4). In detail, 1,931 NCPs were supported by MS and Ribo-seq data, 64 NCPs by MS and ME, 1 NCP by Ribo-seq and ME, and 3 NCPs by MS, Ribo-seq, and ME.

Overview of 9,166 functionally annotated NCPs collected in NCPbook

In NCPbook, we have curated 9,166 functionally annotated NCPs by collecting their functional annotation from 84 peer-reviewed research publications and 4 databases: SPENCER, ncEP, FuncPep, and SmProt (Supplementary Table S5). The sequence information of these functional NCPs is supported by either MS, Ribo-seq, ME, or different combinations of these pieces of evidence (Fig. 2F;

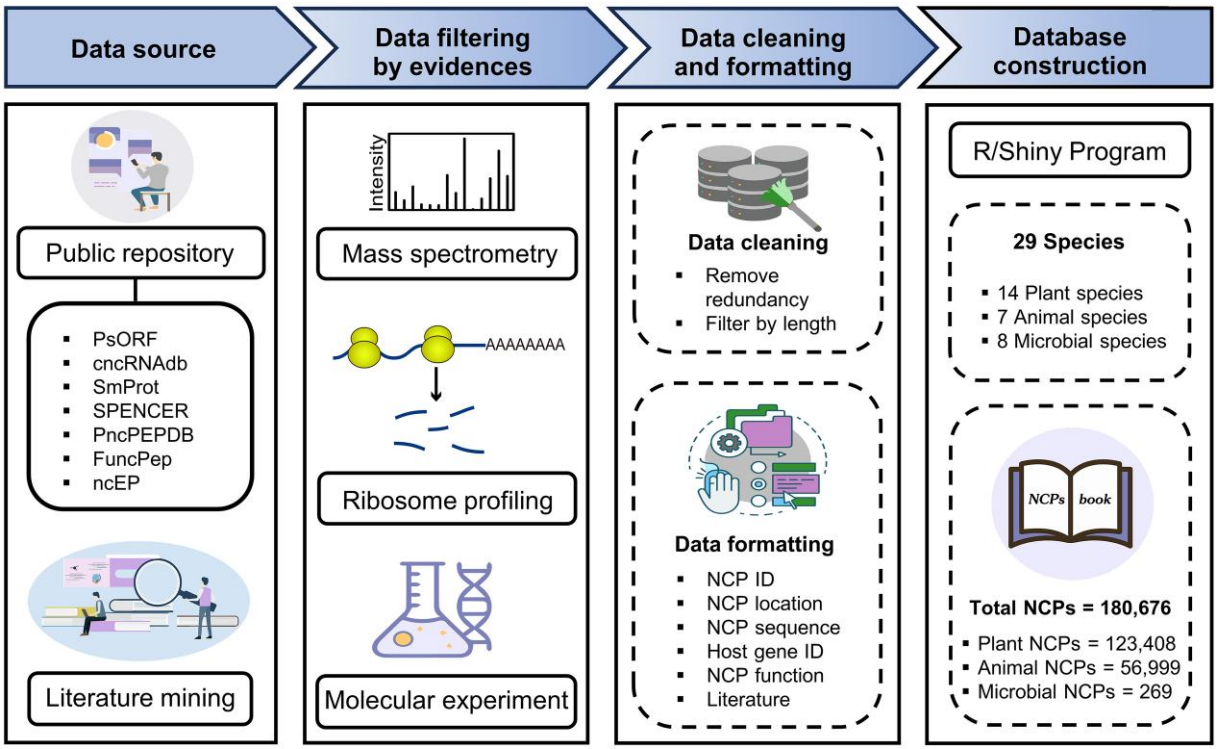

Figure 1. The workflow of data collection and processing to build the NCPbook database. Evidence-supported data were collected from public repositories and research papers. Data cleaning and formatting included the removal of redundant NCPs, assigning NCP ID, NCP genomic location, NCP sequence, host gene ID, NCP function, and referenced literature. The NCPbook was constructed using the R/Shiny software. A total of 180,676 NCPs from 29 species including 14 plant species, 7 animal species, and 8 microbial species were curated in NCPbook. PsORF, plant small ORF; cncRNAdb, coding and ncRNA database; SmProt, small protein; SPENCER, small peptides encoded by ncRNAs in cancer patients; PncPEPDB, putative noncoding peptide database; FuncPep, functional ncRNA-encoded peptides; and ncEP, noncoding experimentally validated peptides.

Supplementary Table S5). Among these, the existence of 2,021 NCPs was evidenced by MS, 7,009 NCPs were evidenced by Ribo-seq, and 92 NCPs were evidenced by ME. The existence of the remaining 44 NCPs was confirmed by the different combinations of MS, Ribo-seq, and ME. In detail, 38 NCPs were evidenced by MS and ME, 1 NCP was evidenced by Ribo-seq and ME, 2 NCPs were evidenced by MS and Ribo-seq, and 3 NCPs were evidenced by MS, Ribo-seq, and ME.

Among the 9,166 functionally annotated NCPs, 9,035 were collected from animals, 104 were collected from plants, and 27 were collected from microbes (Fig. 3). Specifically, the 9,035 NCPs from animals have been categorized into 4 distinct functional categories, comprising of 5,294 NCPs with tumor-specific roles, 2,332 NCPs associated with immunity responses, 352 NCPs involved in inhibitory processes, and 1,057 NCPs with miscellaneous functions. Conversely, the 104 NCPs from plants have been further divided into 4 functional categories: 39 NCPs exhibiting antifungal properties, 24 NCPs playing essential roles in plant growth and development, 3 NCPs associated with cell proliferation, and 38 NCPs contributing to various other processes. Additionally, 27 functionally characterized NCPs with therapeutic, regulatory, and miscellaneous roles from microbial sources were included in the NCPbook (Fig. 3). In summary, NCPbook has collected NCPs with pivotal functional roles across multiple species and kingdoms, highlighting the need for continued efforts to unravel their intricate roles in diverse biological processes.

Functionalities of the NCPbook database

The NCPbook database was constructed using the R/Shiny framework. NCPbook boasts several key attributes designed to enhance its search and analytical capabilities (Fig. 4). The database is structured with 9 distinct modules, denoted as "Home," "Browse," "Search," "BLAST," "JBrowse," "Statistics," "Download," "Help," and "Contact" (Fig. 4). The "Home" page displays an image representing the process used to build the NCPbook database and a paragraph of text introducing NCPs and NCPbook. Under the "Statistics" menu, vital information regarding the source of NCPs and the evidence supporting them is presented. The "Help" menu is further divided into 3 submenus: "About," "Tutorial," and "Links." The "About" submenu gives a brief introduction of the NCPbook database, and the "Tutorial" submenu offers step-by-step guidance on how to navigate and utilize the database effectively. The "Links" submenu displays links of tools and databases related to NCPs. In the "Contact" menu, contact details of the collaborating labs are provided, facilitating direct communication for inquiries regarding the database. The remaining menus within the NCPbook database encompass major analytical functionalities, each of which is elaborated upon in subsequent sections for a more comprehensive understanding.

Browse

The "BROWSE" page in NCPbook provides a convenient means to survey all the collected NCPs. NCPbook features 29 diverse species, including *A. thaliana*, human (*Homo sapiens*), mice (*Mus musculus*), and others. Within the species panel, users are initially directed to select the species group: plant, animal, or microbe that encompasses their specific area of interest. Subsequently, upon selecting a particular species, NCPbook generates an information table presenting all the NCPs collected for the selected species.

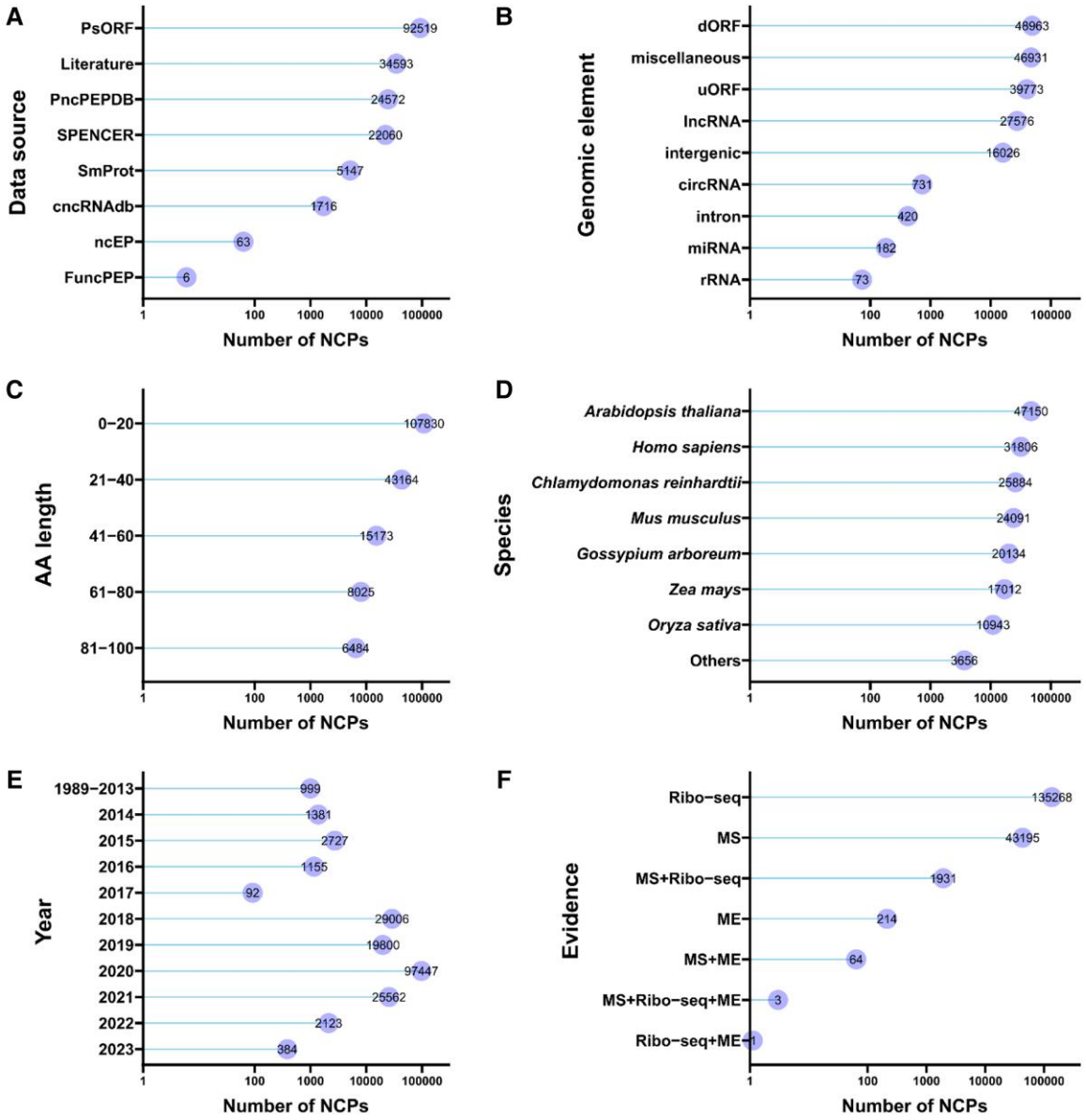

Figure 2. Characteristics of NCPs in NCPbook. **A)** NCPs categorized by data sources. The NCPs were collected from different public databases and published literature. **B)** NCPs categorized by different noncoding regions. **C)** NCPs in NCPbook on the basis of their length. **D)** NCPs categorized by different species of plants, animals, and microbes. **E)** Number of NCPs reported according to their year of publication (1989 to 2023). **F)** Number of reported NCPs evidenced by MS, Ribo-seq, and ME as well as the combination of these supporting evidences. AA, amino acids; MS, mass spectrometry; Ribo-seq, ribosome profiling (Ribo-seq); and ME, molecular experiments.

This table incorporates crucial details such as NCP ID, species name, genomic location, supporting evidence, host gene ID, and type. For further details, users can interact with the information table by clicking on any row, enabling easy access to comprehensive data about the NCPs. This includes NCP ID, sequences, function, length, host gene name, title of the corresponding study, PubMed ID of the study, publish date of the study (year), and database name (Fig. 5A). Additionally, NCPs are cross-referenced with pertinent IDs from external databases, enriching the available information. For all the NCPs retrieved from publicly accessible databases, we have added hyperlinks to the NCP ID, sequence, length, function, and database name in NCPbook. By clicking either of them with hyperlinks, users will be directed to the record page of the corresponding NCP in the original database. Furthermore, there is a search box at the upper-right corner of the table, which enables the users to search for any NCP in a particular species using the NCP ID or other keywords. Moreover, users have the option to easily copy or download the complete NCP information of their desired species by clicking the "copy," "CSV," or "Excel" button located at the upper-left corner of the table.

Search

NCPbook has introduced a "QUICK SEARCH" module on the "Search" page, offering users a direct means to explore the collected NCPs. This module facilitates 3 distinct approaches for querying the database: search by NCP ID, search by host gene ID, and search by genomic location (Fig. 5B). NCP IDs correspond to unique identifiers assigned to each NCP sourced from published literature, databases, as well as those designated by this study. Host gene IDs are the identifiers of the genes from which NCPs originate, while genomic location denotes the specific position of

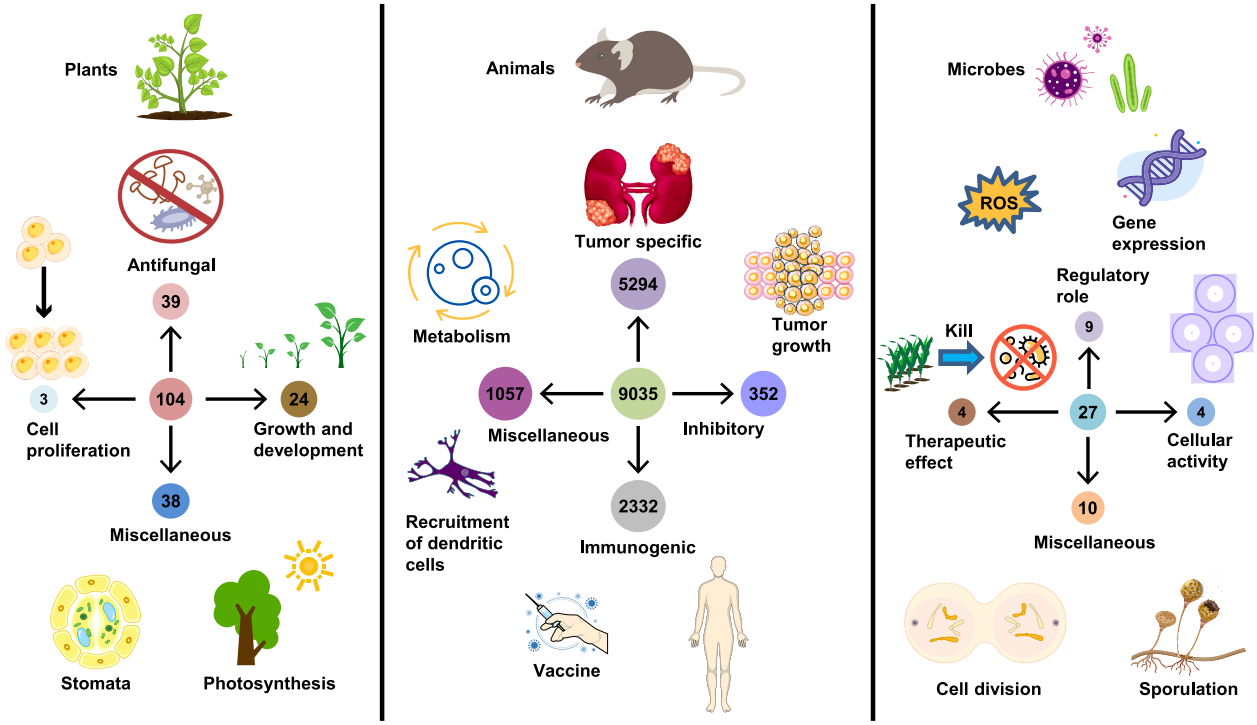

Figure 3. Overview of functionally annotated NCPs collected in NCPbook. The functions include antifungal, cell division, growth and development, nodulation, enzymatic activities, signal transduction, immunotherapy, cell growth, stress response, metabolism, and different regulatory roles in plants, animals, and microbes.

NCPs in the genome. Both host gene IDs and genomic locations are collected from published literature and databases.

To access data regarding input NCPs, users have the option to input 1 or more NCP IDs or host gene IDs across any of the 29 species (Fig. 5B). A concise summary of the search results is presented in tabular format, with each row representing an NCP. Clicking on a specific row within the output table imparts detailed insights into the desired NCPs. This includes comprehensive details such as peptide sequence, length, host gene information, function, supporting experimental evidence, and a description of the corresponding reference. Additionally, the table can be exported as either a CSV or Excel file, enabling users to scrutinize the precise particulars of the NCPs of interest.

BLAST

The BLAST page was designed to facilitate the search for NCPs within 1 or multiple species based on sequence similarity. Users are provided with the option to directly input a query sequence in a FASTA format or upload a FASTA file from their disk. The BLAST page offers 2 distinct blast programs: BLASTP and BLASTX. BLASTX is employed when searching NCPbook with DNA sequences as input, whereas BLASTP is utilized when searching NCPbook with peptide sequences as input. User can choose to conduct BLAST searches against NCPs collected for 1 or multiple species using either default or customized parameters. BLAST searches can then be initiated by clicking the "BLAST" button (Fig. 5C). Upon the completion of the BLAST alignment, users are directed to the output panel under the BLAST menu. In this panel, different BLAST hits detected for the input sequence are displayed in separated rows of a table. Columns of the table display multiple parameters for each BLAST hit, including ID and length of the query sequence, ID and length of the subject sequence, aligned regions between the query and subject, *E*-value, percentage of identical matches, and the species name. For each BLAST hit, more details can be further displayed, by clicking the corresponding row, in the form of a new table encompassing more details such as the bit score, number of gaps, and alignment length. Additionally, sequence alignment between the query (user-input sequence) and the subject (NCP sequence collected in NCPbook) will be displayed on the right of the new table.

JBrowse

The JBrowse page of NCPbook currently offers genome browser services for 8 species, including *A. thaliana*, *Z. mays*, *O. sativa*, *V. vinifera*, *H. sapiens*, *D. melanogaster*, *M. musculus*, and *Escherichia coli*. Upon clicking the image corresponding to a species on the "JBrowse" page, users are directed to the genome browser of the selected species. In JBrowse, 3 distinct tracks are presented, encompassing reference genome sequences, annotated protein-coding gene models, and NCP information. When a specific element within a track is selected, relevant information is presented in the right panel of the genome browser (Fig. 5D).

Download

NCPbook has integrated a "Download" module for users' convenience. Through this module, users can access detailed information and sequences of NCPs collected for all species featured in NCPbook. The files of each species are labeled with its name accompanied by a concise description. To accelerate the downloading process, all files are provided in a gzip format (Fig. 5E).

Case study

miRNA is a well-studied class of ncRNAs playing critical roles in diverse biological processes (Shang et al. 2023). The sequences

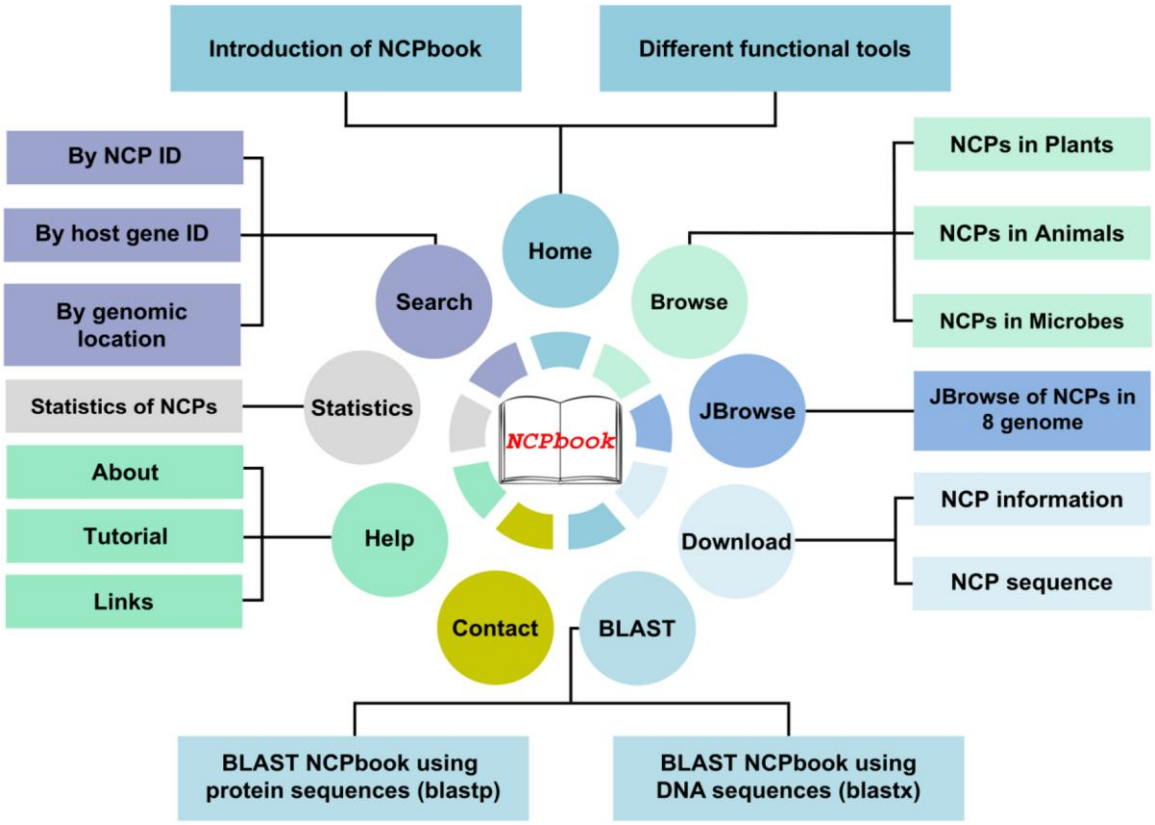

Figure 4. Structure of the NCPbook database. NCPbook is composed of 9 different modules including Home, Browse, Search, BLAST, JBrowse, Statistics, Download, Help, and Contact page.

of miRNAs and their precursors in 271 organisms were comprehensively annotated in miRBase (https://www.mirbase.org/) (Kozomara et al. 2019). Inspired by the vital function of NCPs derived from miRNA precursors in *Arabidopsis* and *V. vinifera* (Lauressergues et al. 2015; Chen, Deng, et al. 2020), we aimed to identify the candidate NCPs derived from miRNAs by aligning the sequences of 38,589 hairpin precursor miRNAs deposited in miRBase against all NCPs collected in NCPbook. By querying the presence of complete NCP sequence in the alignment without gaps or mismatches, we identified 1,045 candidate NCPs derived from miRNAs (Supplementary Table S6). We found that the precursor of miR169a in *Arabidopsis lyrata*, *Brassica napus*, and *Camelina sativa* can also be translated into a peptide with the same sequence as miPEP169a in *A. thaliana* (Supplementary Table S6). Ga08g02251_4 is an NCP associated with a protein-coding gene Ga08g02251 in *G. arboreum*. Ga08g02251_4 (MCSLSS) was aligned with the sequences of 8 miRNA precursors from 7 species without gaps or mismatches (Supplementary Table S6). These findings indicate that a portion of miRNA precursor-encoded NCPs are conserved among different species, which can be selected for further functional exploration in the future.

Discussion

NCPs, as novel biomolecules in plants and animals, are now gaining considerable attention in biological research. This aims to understand their roles and significance across different organisms, opening intriguing avenues for deciphering the mysteries of life (Wilhelm et al. 2014; Zhu et al. 2018; Wang, Yang, et al. 2019; Chen, Deng, et al. 2020; Pei et al. 2022). In the past, NCPs received relatively limited attention due to their origin from previously considered nontranslatable (noncoding) regions of the genome and the lack of accurate approaches for their identification. This is evident from the relatively small number of reported NCPs prior to 2013 (Fig. 2E). Recent advances in Ribo-seq have enabled the high-throughput identification of NCPs (Ingolia et al. 2014; Friedman et al. 2017; Van Heesch et al. 2019). In addition, a strategy known as peptidogenomics has emerged as an effective approach that integrates high-throughput tandem MS and genomic information for NCP identification (Wang et al. 2020). However, computational and experimental problems, including incomplete peptide databases, nonspecific protease digestion, and endogenous peptide enrichment are also major impediments in the identification of plant NCPs. Apart from this, the functions of the most identified plant NCPs are still not known, and the function mechanism of NCPs is far from understanding. Only 104 plant NCPs in the NCPbook were functionally annotated probably caused by the relatively limited attention in the past and the lack of functional studies on NCPs in plants compared with those in human and animals. A comprehensive database of evidence-supported NCPs will be invaluable for future method development in NCP identification.

Biological databases serve as important repositories that play a vital role in facilitating scientists to scrutinize and elucidate various biological phenomena including but not limited to growth, development, stress responses, and organismal metabolism. In the past, considerable efforts were made to build databases housing a vast array of peptides derived from coding regions. For example, PlantPepDB, a plant-derived peptide database, contains 3,848 entries, including both validated and predicted peptides

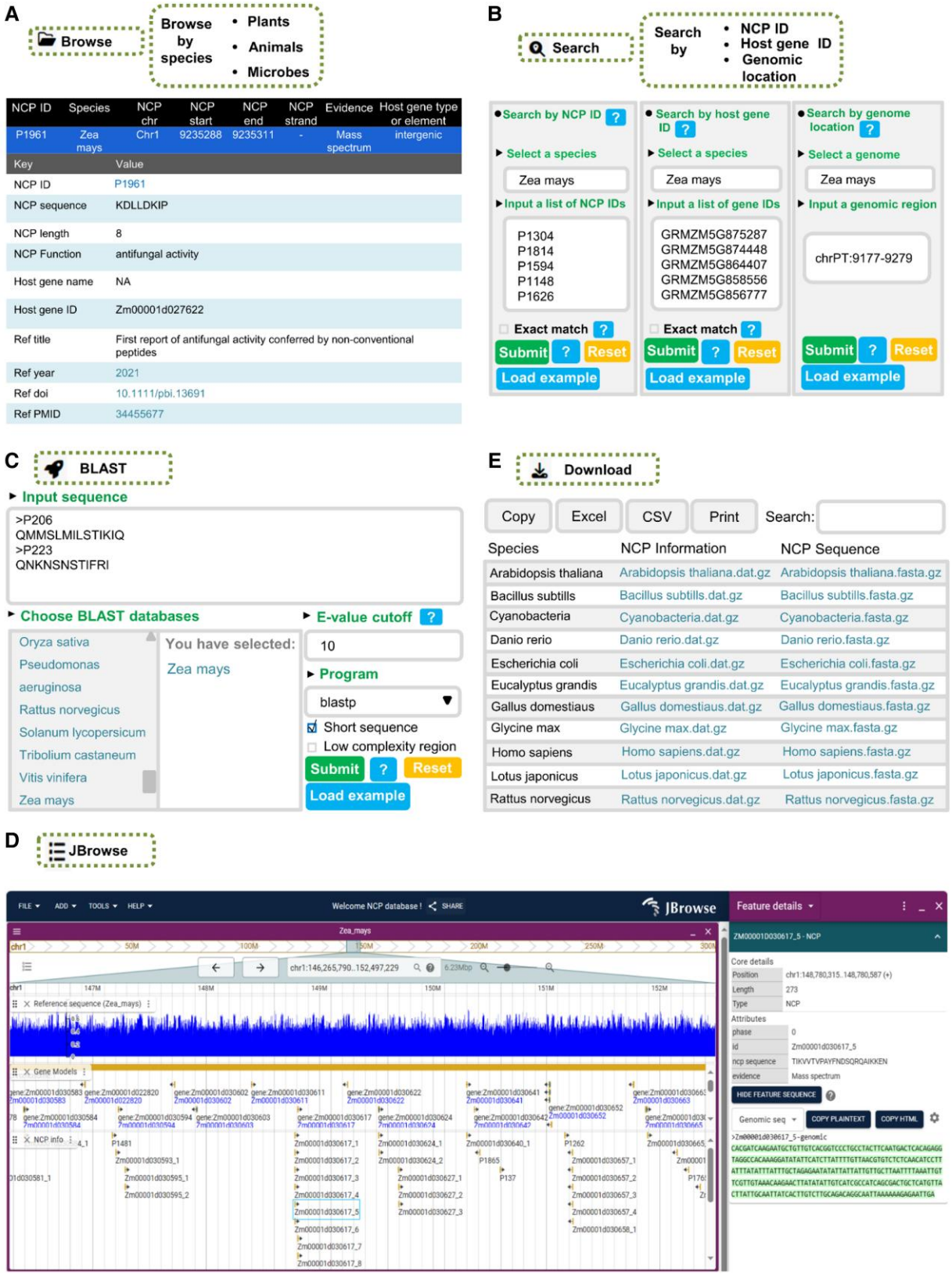

Figure 5. The web interface of NCPbook. **A)** The Browse interface of NCPbook for browsing NCPs in plants, animals, and microbes. The *Z. mays* NCP (P1961) derived from the intergenic region of chromosome 1 is shown in the Browse interface. **B)** The Search interface in NCPbook. NCPs from any 1 of the 29 species can be searched by 1 or multiple NCP IDs, the host gene IDs, or the genomic location. **C)** The BLAST interface for sequence similarity searching in NCPbook. **D)** The JBrowse module of NCPbook. A genomic region of *Z. mays* is displayed with 3 tracks including the reference genome sequence, annotated gene models, and NCPs. **E)** The Download interface of NCPbook.

(Das et al. 2020). MFPPDB reports single or multifunctional therapeutic peptides (1,482,409) from plants (Yang et al. 2023). MtSSPdb is a database of *M. truncatula*, incorporating small secreted peptides with different putative functions (Boschiero et al. 2020). Recently, databases specifically dedicated to NCPs have emerged. For instance, FuncPEP and ncEP incorporate 112 and 74 functional NCPs encoded by ncRNAs, respectively (Dragomir et al. 2020; Liu et al. 2020). LncPep predominantly houses predicted lncRNA-encoded NCPs from public databases (Liu et al. 2022). PncPEPDB and SPENCER contain ncRNA-encoded NCPs supported by Ribo-seq and MS data, respectively (Li et al. 2018; Luo et al. 2022). PsORF comprises sORF-encoded NCPs, including predicted ones (Chen, Li, et al. 2020). SmProt includes ncRNA-encoded NCPs supported by Ribo-seq from 8 species (Li et al. 2021). Despite the valuable information provided by these databases, given the increasing number of identified NCPs and the growing interest in NCPs, there remains a need for a comprehensive NCP database to support modern biological research.

To address this gap, we present NCPbook (https://ncp.wiki/ncpbook/), a comprehensive collection of evidence-supported NCPs in 29 eukaryotic and prokaryotic genomes. NCPbook was constructed by integrating 180,676 NCPs from published literatures and multiple publicly accessible databases. This database houses a vast array of NCPs derived from a variety of noncoding gene elements (introns, intergenic regions, uORFs, dORFs, lncRNAs, miRNAs, circRNAs, and rRNAs) across a diverse range of species. NCPbook offers a range of functionalities allowing users to browse, search, perform sequence alignments (BLAST), and visualize genomic data (JBrowse) related to NCPs. Moreover, users can download the full datasets deposited in NCPbook, which will empower researchers to explore the "coding realm" of these various noncoding regions across diverse biological systems. A "Help" page is offered for users to understand and use this database efficiently. NCPbook could serve as a useful platform for researchers in comparative genomics and functional studies. For instance, NCPbook can be utilized in genome-wide association studies and quantitative trait loci mapping to investigate the presence of NCPs in specific genomic regions significantly associated with phenotypic variations. We further demonstrate the utility of NCPbook in identifying candidate NCPs derived from miRNA precursors by aligning sequences from miRBase against NCPs collected in NCPbook. This alignment led to the identification of 1,045 candidate NCPs, including NCPs with identical peptide sequences to miPEP169a in *A. thaliana* and Ga08g02251_4 in *G. arboreum* (Supplementary Table S6). Our findings also suggest that some miRNA-associated NCPs are conserved across different species, highlighting their potential for future functional exploration.

In summary, NCPbook offers a user-friendly interface and a range of analytical tools that streamline the investigation of NCPs, providing researchers with easy access to a wealth of information. As the field of NCP research continues to expand, NCPbook will incorporate new discoveries. This involves integrating additional NCPs supported by MS, Ribo-seq, or ME. We will regularly monitor the latest literature using monthly alerts in PubMed, Google Scholar, and Web of Science with relevant keywords. Furthermore, strengthening the connection with other databases, such as UniProt (The UniProt Consortium 2023), Gramene (Marcela et al. 2022), and miRbase (Kozomara et al. 2019), would enrich the user experience and enable comprehensive cross-referencing of information related to NCPs and their host genes. In all, NCPbook will facilitate and advance research in the growing field of NCPs, contributing to a better understanding of the roles played by NCPs in biological processes across diverse organisms.

Materials and methods

Data collection

The NCPs in NCPbook were curated from peer-reviewed literature and multiple publicly accessible databases. To initiate this comprehensive collection, we conducted keyword searches employing terms such as "non-canonical peptides," "non-conventional peptides," "ncRNA-encoded peptides," "lncRNA-encoded peptides," "circRNA-encoded peptides," and "miRNA-encoded peptides" across prominent scholarly platforms, including Google Scholar, ScienceDirect, Web of Science, and PubMed. Subsequently, all identified literature underwent a preliminary evaluation conducted by experienced curators to eliminate unrelated articles. Additionally, to avoid omitting any newly discovered NCPs, we double-checked all the papers that cited at least 1 of the remaining pertinent articles.

Data processing

A standard pipeline was used to process all of the gathered data. Data extraction from various public databases was performed using an in-house R code. The NCPbook database encompasses several key pieces of information, including the peptide's name assigned by the authors, the associated literature, its symbol, and the name assigned by NCBI. Furthermore, we manually extracted data from diverse sources, incorporating details concerning ncRNAs (including the type of ncRNA and its ID), peptides (name, sequence, and length), experimental particulars (species and techniques employed), peptide functions, and entry sources (paper title, PubMed ID, description, etc.). Subsequently, we conducted a reannotation of essential information, encompassing peptide genome locations, peptide sequences, and noncoding genomic positions (Fig. 1). To ensure uniform formatting, all NCP sequences were stored in FASTA format, and the associated information was stored in plain text files. The process of data normalization involved combining all the standardized data files using the R code.

Construction of the NCPbook database

The NCPbook database was built using the R/Shiny framework, a widely recognized tool for creating web applications in data science (R Core team 2023; Yu et al. 2019; Chang et al. 2021; Jia et al. 2022). In essence, the NCPbook application comprises 2 fundamental components: server.R and ui.R. Specifically, ui.R defines the graphical interface of NCPbook, facilitating user interactions such as parameter configuration and data uploading. By converting the R code into JavaScript, CSS, and HTML, ui.R governs the layout and visual aspects of NCPbook. Additionally, ui.R is responsible for generating HTML widgets, including action buttons, download buttons, slider inputs, radio buttons, and checkboxes, which serve as the primary building blocks of NCPbook. All user inputs processed through ui.R are monitored by server.R, which subsequently coordinates the computational processes and displays the results within the NCPbook graphical user interface.

Acknowledgments

We thank Dr. Hao Guo for his kind suggestions for website art design.

Author contributions

L.W. conceived the original research plan and supervised the project. W.Y. provided bioinformatic platform for the construction of

NCPbook. A.S., W.Y., U.A., X.C., L.T., S.W., M.C., J.Z., and H.L. collected the raw data. A.S., H.Y., and M.F. processed the data. M.F., W.Y., A.S., and H.Y. built the database. A.S. wrote the manuscript. L.W. and W.Y. revised the manuscript. All the authors contributed to the article and approved the manuscript.

Supplementary data

The following materials are available in the online version of this article.

Supplementary Table S1. Different plants, animals, and microbial sources of NCPs.

Supplementary Table S2. Number of NCPs retrieved from published articles and databases.

Supplementary Table S3. Plants, animals, and microbial species retrieved from databases and published literature.

Supplementary Table S4. Collection of evidence-supported NCPs from plants, animals, and microbes.

Supplementary Table S5. A list of 9,166 functionally annotated NCPs curated in NCPbook.

Supplementary Table S6. Candidate NCPs identified in the hairpin precursor of miRNAs.

Funding

This study was supported by National Key Research and Development Program of China (2022YFD1201802 to L.W.), and National Natural Science Foundation of China (U22A20474 and 32172073 to L.W.).

Conflict of interest statement. None declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

References

Anderson DM, Anderson KM, Chang CL, Makarewich CA, Nelson BR, McAnally JR, Kasaragod P, Shelton JM, Liou J, Bassel-Duby R, *et al.* A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 2015:160(4):595–606. https://doi.org/10.1016/j.cell.2015.01.009

Baxevanis AD, Bateman A. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*. 2015:50(1):1.1.1–1.1.8. https://doi.org/10.1002/0471250953.bi0101s50

Boschiero C, Dai X, Lundquist PK, Roy S, Christian de Bang T, Zhang S, Zhuang Z, Torres-Jerez I, Udvardi MK, Scheible WR, *et al.* MtSSPdb: the *Medicago truncatula* small secreted peptide database. *Plant Physiol*. 2020:183(1):399–413. https://doi.org/10.1104/pp.19.01088

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. shiny: Web application framework for R [Computer software]. R package version 1.6. 0. 2021:7211–7229.

Chen QJ, Deng BH, Gao J, Zhao ZY, Chen ZL, Song SR, Wang L, Zhao LP, Xu WP, Zhang CX, *et al.* A miRNA-encoded small peptide, vvi-miPEP171d1, regulates adventitious root formation. *Plant Physiol*. 2020:183(2):656–670. https://doi.org/10.1104/pp.20.00197

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):2158–2160. https://doi.org/10.1111/pbi.13389

Couzigou M, André O, Guillotin B, Alexandre M, Combier P. Use of microRNA-encoded peptide miPEP172c to stimulate nodulation in soybean. *New Phyto*. 2016:211(2):379–381. https://doi.org/10.1111/nph.13991

Das D, Jaiswal M, Khan FN, Ahamad S, Kumar S. PlantPepDB: a manually curated plant peptide database. *Sci Rep*. 2020:10(1):2194. https://doi.org/10.1038/s41598-020-59165-2

Dragomir MP, Manyam GC, Ott LF, Berland L, Knutsen E, Ivan C, Lipovich L, Broom BM, Calin GA. FuncPEP: a database of functional peptides encoded by non-coding RNAs. *Noncoding RNA* 2020:6(4):41. https://doi.org/10.3390/ncrna6040041

Friedman RC, Kalkhof S, Doppelt-Azeroual O, Mueller SA, Chovancova M, von Bergen M, Schwikowski B. Common and phylogenetically widespread coding for peptides by bacterial small RNAs. *BMC Genomics* 2017:18(1):553. https://doi.org/10.1186/s12864-017-3932-y

Huang Y, Wang J, Zhao Y, Wang H, Liu T, Li Y, Cui T, Li W, Feng Y, Luo J, *et al.* cncRNAdb: a manually curated resource of experimentally supported RNAs with both protein-coding and noncoding function. *Nucleic Acids Res*. 2021:49(D1):D65–D70. https://doi.org/10.1093/nar/gkaa791

Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJ, Jackson SE, Wills MR, Weissman JS. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*. 2014:8(5):1365–1379. https://doi.org/10.1016/j.celrep.2014.07.045

Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, Khan OM, Brewer JR, Skadow MH, Duizer C. The translation of non-canonical open reading frames controls mucosal immunity. *Nature* 2018:564(7736):434–438. https://doi.org/10.1038/s41586-018-0794-7

Jia L, Li Y, Huang F, Jiang Y, Li H, Wang Z, Chen T, Li J, Zhang Z, Yao W. LIRBase: a comprehensive database of long inverted repeats in eukaryotic genomes. *Nucleic Acids Res*. 2022:50(D1):D174–D182. https://doi.org/10.1093/nar/gkab912

Kondo T, Plaza S, Zanet J, Benrabah E, Valenti P, Hashimoto Y, Kobayashi S, Payre F, Kageyama Y. Small peptides switch the transcriptional activity of Shavenbaby during *Drosophila* embryogenesis. *Science* 2010:329(5989):336–339. https://doi.org/10.1126/science.1188158

Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 2019:47(D1):D155–D162. https://doi.org/10.1093/nar/gky1141

Lauressergues D, Couzigou JM, Clemente HS, Martinez Y, Dunand C, Bécard G, Combier JP. Primary transcripts of microRNAs encode regulatory peptides. *Nature* 2015:520(7545):90–93. https://doi.org/10.1038/nature14346

Lauressergues D, Ormancey M, Guillotin B, San Clemente H, Camborde L, Duboé C, Tourneur S, Charpentier P, Barozet A, Jauneau A, *et al.* Characterization of plant microRNA-encoded peptides (miPEPs) reveals molecular mechanisms from the translation to activity and specificity. *Cell Rep*. 2022:38(6):110339. https://doi.org/10.1016/j.celrep.2022.110339

Li Q, Ahsan MA, Chen H, Xue J, Chen M. Discovering putative peptides encoded from noncoding RNAs in ribosome profiling data of *Arabidopsis thaliana*. *ACS Synth Biol*. 2018:7(2):655–663. https://doi.org/10.1021/acssynbio.7b00386

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 2021:19(4):602–610. https://doi.org/10.1016/j.gpb.2021.09.002

Liu H, Zhou X, Yuan M, Zhou S, Huang YE, Hou F, Song X, Wang L, Jiang W. ncEP: a manually curated database for experimentally validated ncRNA-encoded proteins or peptides. *J Mol Biol*. 2020:432(11):3364–3368. https://doi.org/10.1016/j.jmb.2020.02.022

Liu T, Wu J, Wu Y, Hu W, Fang Z, Wang Z, Jiang C, Li S. LncPep: a resource of translational evidences for lncRNAs. *Front Cell Dev Biol*. 2022:10:795084. https://doi.org/10.3389/fcell.2022.795084

Luo X, Huang Olexiouk Y, Li H, Luo Y, Zuo Z, Ren J, Xie Y. SPENCER: a comprehensive database for small peptides encoded by noncoding RNAs in cancer patients. *Nucleic Acids Res*. 2022:50(D1):D1373–D1381. https://doi.org/10.1093/nar/gkab822

Marcela KT, Pankaj J, Doreen W. Gramene: a resource for comparative analysis of plants genomes and pathways. *Methods Mol Biol*. 2022:2443:101–131. https://doi.org/10.1007/978-1-0716-2067-0_5

Ormancey M, Guillotin B, San Clemente H, Thuleau P, Plaza S, Combier JP. Use of microRNA-encoded peptides to improve agronomic traits. *Plant Biotechnol J*. 2021:19(9):1687–1689. https://doi.org/10.1111/pbi.13654

Pei MS, Liu HN, Wei TL, Yu YH, Guo DL. Large-scale discovery of non-conventional peptides in grape (*Vitis vinifera* L.) through peptidogenomics. *Hortic Res*. 2022:9:uhac023. https://doi.org/10.1093/hr/uhac023

R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2023. https://www.R-project.org/

Shang R, Lee S, Senavirathne G, Lai E. microRNAs in action: biogenesis, function and regulation. *Nat Rev Genet*. 2023:24(12):816–833. https://doi.org/10.1038/s41576-023-00611-y

Sharma A, Badola K, Bhatia C, Sharma D, Trivedi K. Primary transcript of miR858 encodes regulatory peptide and controls flavonoid biosynthesis and development in *Arabidopsis*. *Nat Plants*. 2020:6(10):1262–1274. https://doi.org/10.1038/s41477-020-00769-x

Sun L, Wang W, Han C, Huang W, Sun Y, Fang K, Zeng Z, Yang Q, Pan Q, Chen T, *et al*. The oncomicropeptide APPLE promotes hematopoietic malignancy by enhancing translation initiation. *Mol Cell*. 2021:81(21):4493–4508. https://doi.org/10.1016/j.molcel.2021.08.033

Tavormina P, De Coninck B, Nikonorova N, De Smet I, Cammue BP. The plant peptidome: an expanding repertoire of structural features and biological functions. *Plant Cell* 2015:27(8):2095–2118. https://doi.org/10.1105/tpc.15.00440

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 2023:51(D1):D523–D531. https://doi.org/10.1093/nar/gkac1052

Tian L, Chen X, Jia X, Wang S, Wang X, Zhang J, Zhang Y, Wu S, Chen Y, Wu L. First report of antifungal activity conferred by non-conventional peptides. *Plant Biotechnol J*. 2021:19(11):2147–2149. https://doi.org/10.1111/pbi.13691

Vale M, Rodrigues J, Badim H, Gerós H, Conde A. Exogenous application of non-mature miRNA-encoded miPEP164c inhibits proanthocyanidin synthesis and stimulates anthocyanin accumulation in grape berry cells. *Front Plant Sci*. 2021:12:706679. https://doi.org/10.3389/fpls.2021.706679

Van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, Kirchner M, Maatz H, Blachut S, Sandmann CL, *et al*. The translational landscape of the human heart. *Cell* 2019:178(1):242–260. https://doi.org/10.1016/j.cell.2019.05.010

Wang S, Mao C, Liu S. Peptides encoded by noncoding genes: challenges and perspectives. *Signal Transduct Target Ther*. 2019:4(1):57. https://doi.org/10.1038/s41392-019-0092-3

Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2. 0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res*. 2019:47(D1):D230–D234. https://doi.org/10.1093/nar/gky978

Wang S, Tian L, Liu H, Li X, Zhang J, Chen X, Jia X, Zheng X, Wu S, Chen Y, *et al*. Large-scale discovery of non-conventional peptides in maize and *Arabidopsis* through an integrated peptidogenomic pipeline. *Mol Plant*. 2020:13(7):1078–1093. https://doi.org/10.1016/j.molp.2020.05.012

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, *et al*. Mass-spectrometry-based draft of the human proteome. *Nature* 2014:509(7502):582–587. https://doi.org/10.1038/nature13319

Xing J, Liu H, Jiang W, Wang L. LncRNA-encoded peptide: functions and predicting methods. *Front Oncol*. 2021:10:622294. https://doi.org/10.3389/fonc.2020.622294

Yang Y, Wu H, Gao Y, Tong W, Li K. MFPPDB: a comprehensive multifunctional plant peptide database. *Front Plant Sci*. 2023:14:1224394. https://doi.org/10.3389/fpls.2023.1224394

Yu Y, Yao W, Wang Y, Huang F. ShinyChromosome: an R/shiny application for interactive creation of non-circular plots of whole genomes. *Genomics Proteomics Bioinformatics* 2019:17(5):535–539. https://doi.org/10.1016/j.gpb.2019.07.003

Zhu S, Wang J, He Y, Meng N, Yan GR. Peptides/proteins encoded by non-coding RNA: a novel resource bank for drug targets and biomarkers. *Front Pharmacol*. 2018:9:1295. https://doi.org/10.3389/fphar.2018.01295