CCPRD: A Novel Analytical Framework for the Comprehensive Proteomic Reference Database Construction of NonModel Organisms

Qingxiang Guo, Dan Li, Yanhua Zhai, and Zemao Gu*

ABSTRACT: Protein reference databases are a critical part of producing efficient proteomic analyses. However, the method for constructing clean, efficient, and comprehensive protein reference databases of nonmodel organisms is lacking. Existing methods either do not have contamination control procedures, or these methods rely on a three-frame and/or six-frame translation that sharply increases the search space and the need for computational resources. Herein, we propose a framework for constructing a customized comprehensive proteomic reference database (CCPRD) from draft genomes and deep sequencing transcriptomes. Its effectiveness is demonstrated by incorporating the proteomes of nematocysts from endoparasitic cnidarian: myxozoans. By applying customized contamination removal procedures, contaminations in omic data were successfully identified and removed. This is an effective method that does not result in overdecontamination. This can be shown by comparing the CCPRD MS results with an artificially contaminated database and another database with removed contaminations in genomes and transcriptomes added back. CCPRD outperformed traditional frame-based methods by identifying 35.2−50.7% more peptides and 35.8−43.8% more proteins, with a maximum of 84.6% in size reduction. A BUSCO analysis showed that the CCPRD maintained a relatively high level of completeness compared to traditional methods. These results confirm the superiority of the CCPRD over existing methods in peptide and protein identification numbers, database size, and completeness. By providing a general framework for generating the reference database, the CCPRD, which does not need a high-quality genome, can potentially be applied to nonmodel organisms and significantly contribute to proteomic research.

INTRODUCTION

Recent advances in proteomics have improved our understanding of the molecular events that underpin the development, evolution, and the ecology of life.1−4 In those proteomic experiments, the identification of peptide spectra depends on searching a spectrum that is matched in the respective protein sequence databases. These databases are usually derived from the reference genome of a particular organism and are critical for downstream analyses.5 However, constructing an efficient and comprehensive protein reference database can be fraught with challenges because obtaining high-quality genomes for new organisms is still time-intensive and expensive today. This is especially true for species with a polyploidy nature or abnormal GC content.6−8 Therefore, most early attempts to cope with the proteomic reference databases have focused on public databases such as the NCBI nr and the UniProtKB/Swiss-Prot, or used the proteins predicted from relative species.9−11 These methods are either computationally expensive or have a poor level of completeness.12 Later on, this situation was improved by the addition of RNA-seq data that were thought to offer many advantages such as reduced cost and a direct reflection of coding sequences. However, together with genomic data, these RNA-seq results were usually processed by three-frame and six-frame translations before being sent to a reference database construction.13−16 Most proteins obtained through this method are in reading frames that are not translated, unnecessarily increasing the search space for spectral matching. Even though modern search strategies like Prosi17 can overcome the FDR increase in the larger databases, big search space still requires more computational resources.18 Recently, several good approaches have been developed to provide ways beyond classic 6-frame translation-based methods, such as MSProGene,19 JUMPg,20 and ProteomeGenerator.21 These methods are effective and focused on aspects such as search space reduction, gene annotation improvement, and isoform identification. However, contamination can be introduced during processes such as sampling and
culturing, and few studies have been devoted to the production of protein reference sets that are free of contamination in the context of proteomics. This requires a delicate balance between decontamination and overcontamination. As a result, a method for constructing a clean, efficient, and comprehensive proteomic reference database for those species with heavy contaminations and little knowledge about translated proteins is desired.

Myxozoans are obligate endoparasitic cnidarians that mainly infect fish. This lineage is known as the most basal metazoan parasite and achieves vast success in biodiversity, representing an intriguing riddle for evolutionary biologists, ecologists, and a still underexploited model for proteomics. In fact, myxozoans have rarely been used for MS with a holistic consideration of reference database and contamination control, likely because of their nature. Only a few species can be experimentally infected in a lab, and most sampling processes are highly dependent on the season and contain contamination.

The aim of this study was to provide methods for constructing a clean, efficient, and comprehensive proteomic reference database, referred to as a customized comprehensive proteomic reference database (CCPRD). Its effectiveness was tested with the nematocysts of myxozoans, representing an extreme case that involved micrometer-sized organelles isolated from organisms that contained heavy contamination. To populate the database, we generated draft genomes and a deep coverage transcriptome (RNA-seq) of three myxozoans and screened them against host and bacterial sequences used for removing contamination. Then, using the knowledge of reading frame detection, frameshift correction, and annotation, the predicted proteins were incorporated into the CCPRD. The
performance of the CCPRD was assessed by comparing its MS results with four alternative databases: (1) a six-frame translation of the transcriptomes (trans_6_frame); (2) a six-frame translation of the transcriptome plus genomes (genome +transcriptome_6_frame); (3) a contaminant database made by adding manually selected host and bacteria sequences to the CCPRD (CCPRD_contam); and (4) adding back sequences (both from genomes and transcriptomes) that were removed by the decontamination process to the CCPRD (CCPRD_remove). All of the databases were evaluated with respect to size, the number of identified peptides and proteins, and completeness. In proteomic experiments of myxozoan nematoxysts, our database significantly outperforms alternative approaches in peptide and protein identification numbers, database size, and completeness. The CCPRD facilitated the proteomic analysis by identifying 19.1–43.8% more proteins with a maximum of 84.6% in database size reduction. This is by far the most comprehensive proteomic study on myxozoans, and one of the deepest proteomes predicted on any endoparasites. Last, we offer the means and the scripts for the CCPRD construction of their own proteomic experiments on any organism.

Figure 3. Blob plot of the initial M. honghuensis Liu et Gu, 2011 genome assembly (a, left panel) and the initial M. wulii Landsberg & Lom genome assembly (b, left panel), displaying bacteria contamination and host contamination. Each scaffold is drawn as a circle based on its GC content (X-axis) and coverage (Y-axis), with a diameter proportional to its length and colored by its taxonomic annotation at the phylum-level. In the legend, colors of phyla are listed together with scaffold-count, scaffold span, and scaffold N50. The histograms above and to the right of the main scatter plot sum contig spans for GC proportion bins and coverage bins, respectively. Blob plot of the M. honghuensis assembly (a, right panel) and M. wulii assembly (b, right panel) after the removal of contaminant scaffolds.
RESULTS

CCPRD and Alternative Databases for the Proteomics of Myxozoan Nematocysts. An overview of the process we used to generate and evaluate our reference database is shown in Figure 1.

The CCPRD was then applied to the proteomic analysis of myxozoan nematocysts, the workflow of which is shown in Figure 2.

The CCPRD framework requires an initial data set of low-depth genomes and RNA-sequencing. For this purpose, a total of 218.1, 100.7, and 218.6 Mb paired-end (PE) transcriptomic data were generated for *Myxobolus honghuensis*, *Myxobolus wulii*, and *Thelohanellus kitauei*. For *M. honghuensis* and *M. wulii*, 160.1 and 64.8 Mb PE genomic data were produced, respectively. The detailed analyses of the genomes and transcriptomes will be published elsewhere (Guo et al., article in preparation). After initial assembly, 27,502, 42,948, and 90,775 unigenes were generated for *M. honghuensis*, *M. wulii*, and *T. kitauei*, respectively. For *M. honghuensis* and *M. wulii*, 295 and 235 Mb draft genomes were assembled, respectively. Upon removal of the sequences assigned to the host or bacterial origin (Table S1), 26,886, 42,380, and 88,081 unigenes were retained for these. For the *M. honghuensis* genome, 16.05 Mb *Proteobacteria*, 1.99 Mb *Chordata*, and 5.27 Mb *Firmicutes* sequences were removed. For the *M. wulii* genome, 0.8 Mb *Streptophyta* and 2.51 Mb *Chordata* sequences were removed (Figure 3).

Cleaned assemblies were then assembled into five databases, as described previously. A summary of the composition of CCPRD and alternative databases is shown in Table 1.

| species       | database name         | database components                                                                 | sequence tag | sequence number | database size |
|---------------|-----------------------|-------------------------------------------------------------------------------------|--------------|-----------------|---------------|
| *M. honghuensis* | H_CCPRD               | CCPRD, including informations from genome and transcriptome                          | MH_MC        | 173,004         | 21.2 Mb       |
|               | H_trans_6_frame       | transcriptome six-frame translation                                                  | MH_T6        | 186,638         | 13.1 Mb       |
|               | H_genome+transcriptome_6_frame | genome and transcriptome six-frame translation                                    | MH_A6        | 1,806,584      | 100 Mb        |
|               | H_CCPRD_contam        | CCPRD + contaminants                                                                  | MH_MC/HOST/BAC| 278,298         | 66.5 Mb       |
|               | H_CCPRD_remove        | CCPRD + sequences removed in decontamination process                                 | MH_MC/RM     | 429,754         | 37.7 Mb       |
| *M. wulii*    | W_CCPRD               | CCPRD, including information from genome and transcriptome                            | WL_MC        | 170,351         | 25.7 Mb       |
|               | W_trans_6_frame       | transcriptome six-frame translation                                                  | WL_T6        | 254,221         | 21.2 Mb       |
|               | W_genome+transcriptome_6_frame | genome and transcriptome six-frame translation                                    | WL_A6        | 1,795,435       | 106 Mb        |
|               | W_CCPRD_contam        | CCPRD + contaminants                                                                  | WL_MC/HOST/BAC| 275,645         | 70.9 Mb       |
|               | W_CCPRD_remove        | CCPRD + sequences removed in decontamination process                                 | WL_MC/RM     | 216,320         | 28.3 Mb       |
| *T. kitauei*  | T_CCPRD               | CCPRD, including information from genome and transcriptome                            | TK_MC        | 184,002         | 20.9 Mb       |
|               | T_trans_6_frame       | transcriptome six-frame translation                                                  | TK_T6        | 546,831         | 37.3 Mb       |
|               | T_genome+transcriptome_6_frame | genome and transcriptome six-frame translation                                    | TK_A6        | 2,209,730       | 135 Mb        |
|               | T_CCPRD_contam        | CCPRD + contaminants                                                                  | TK_MC/HOST/BAC| 289,296         | 66.1 Mb       |
|               | T_CCPRD_remove        | CCPRD + sequences removed in decontamination process                                 | TK_MC/RM     | 196,635         | 21.7 Mb       |

**CCPRD**: customized comprehensive proteomic reference database.

| species       | database name         | database components                                                                 | sequence tag | sequence number | database size |
|---------------|-----------------------|-------------------------------------------------------------------------------------|--------------|-----------------|---------------|
| *M. honghuensis* | H_CCPRD               | CCPRD, including informations from genome and transcriptome                          | MH_MC        | 173,004         | 21.2 Mb       |
|               | H_trans_6_frame       | transcriptome six-frame translation                                                  | MH_T6        | 186,638         | 13.1 Mb       |
|               | H_genome+transcriptome_6_frame | genome and transcriptome six-frame translation                                    | MH_A6        | 1,806,584      | 100 Mb        |
|               | H_CCPRD_contam        | CCPRD + contaminants                                                                  | MH_MC/HOST/BAC| 278,298         | 66.5 Mb       |
|               | H_CCPRD_remove        | CCPRD + sequences removed in decontamination process                                 | MH_MC/RM     | 429,754         | 37.7 Mb       |
| *M. wulii*    | W_CCPRD               | CCPRD, including information from genome and transcriptome                            | WL_MC        | 170,351         | 25.7 Mb       |
|               | W_trans_6_frame       | transcriptome six-frame translation                                                  | WL_T6        | 254,221         | 21.2 Mb       |
|               | W_genome+transcriptome_6_frame | genome and transcriptome six-frame translation                                    | WL_A6        | 1,795,435       | 106 Mb        |
|               | W_CCPRD_contam        | CCPRD + contaminants                                                                  | WL_MC/HOST/BAC| 275,645         | 70.9 Mb       |
|               | W_CCPRD_remove        | CCPRD + sequences removed in decontamination process                                 | WL_MC/RM     | 216,320         | 28.3 Mb       |
| *T. kitauei*  | T_CCPRD               | CCPRD, including information from genome and transcriptome                            | TK_MC        | 184,002         | 20.9 Mb       |
|               | T_trans_6_frame       | transcriptome six-frame translation                                                  | TK_T6        | 546,831         | 37.3 Mb       |
|               | T_genome+transcriptome_6_frame | genome and transcriptome six-frame translation                                    | TK_A6        | 2,209,730       | 135 Mb        |
|               | T_CCPRD_contam        | CCPRD + contaminants                                                                  | TK_MC/HOST/BAC| 289,296         | 66.1 Mb       |
|               | T_CCPRD_remove        | CCPRD + sequences removed in decontamination process                                 | TK_MC/RM     | 196,635         | 21.7 Mb       |

**Table 1. Information of Different Databases for MS/MS Search**

**Table 2. Information of Differently Databases for MS/MS Search**

(a) Comparison of the protein reference databases in the number of identified unique peptides (a) and proteins (b) for *M. honghuensis* Liu et Gu, 2011, *M. wulii* Landsberg & Lom, 1991 and *T. kitauei* Egusa & Nakajima, 1981 nematocyst samples. The number of unique peptides or proteins was identified with 1% FDR. CCPRD significantly outperforms the traditional database construction method. Trans_6_frame, transcriptome six-frame translation; genome+transcriptome_6_frame, genome, and transcriptome six-frame translation; CCPRD_contam, CCPRD + contaminants from host and bacteria; CCPRD_remove, CCPRD + sequences removed in the decontamination process. See detailed information on these databases in Table 1.

![Figure 4](https://dx.doi.org/10.1021/acsomega.0c01278)
The largest database size (100−135 Mb) and is approximately 4.1−6.5 times the size of the respective CCPRD. For all species, the CCPRD has the lowest sequence number (173,004−184,002) compared to the other four databases, and only constituted 8.32−9.57% of the largest genome+transcriptome_6_frame databases.

**Database Evaluation by the Number of Identified Peptides and Proteins.** For all three species, the CCPRD enabled the identification of more peptides and proteins compared with the databases trans_6_frame, genome+transcriptome_6_frame, and CCPRD_remove, respectively. The CCPRD_contam identified the most peptides in *M. honghuensis* and *T. kitauei*, while the CCPRD outperformed all the others in *M. wulii*. Adding bacteria and creating host contamination resulted in the CCPRD_contam identifying 15.9 and 10.1% more peptides than the CCPRD in *M. honghuensis* and *T. kitauei*, respectively, except that with *M. wulii* the CCPRD identified 4.7% more peptides than CCPRD_contam. Besides, in all three species, the trans_6_frame identified 15.8, 22.3, and 11.5% more peptides than the database genome+transcriptome_6_frame, which identified the least peptides among the five databases.

The comparison results in the number of identified proteins are generally in accordance with those of peptides. For all species, the CCPRD identified 16.2−35.8, 19.1−43.8, and 1.5−

Table 2. Composition Details from Identification Results of CCPRD, CCPRD + Host and Bacteria Sequences (CCPRD_contam), and CCPRD + Sequences Removed in the Decontamination Process (CCPRD_remove) of *M. honghuensis*, *M. wulii*, and *T. kitauei*.

| database name                  | MYXO | MYXO | HOST | BAC | total | MYXO | RM | total |
|-------------------------------|------|------|------|-----|-------|------|----|-------|
| peptide results               |      |      |      |     |       |      |    |       |
| *M. honghuensis*              | 3411 | 2999 | 461  | 493 | 3953  | 3060 | 102| 3162  |
| *M. wulii*                    | 1963 | 1545 | 115  | 214 | 1874  | 1863 | 22 | 1885  |
| *T. kitauei*                  | 1913 | 1406 | 511  | 189 | 2106  | 1879 | 16 | 1895  |
| protein results               |      |      |      |     |       |      |    |       |
| *M. honghuensis*              | 1070 | 1008 | 164  | 260 | 1432  | 1004 | 50 | 1054  |
| *M. wulii*                    | 511  | 412  | 29   | 114 | 555   | 490  | 4  | 494   |
| *T. kitauei*                  | 593  | 440  | 175  | 108 | 723   | 562  | 9  | 571   |

**MYXO: myxozoans, BAC: bacterial sequences, RM: sequences removed in the decontamination process.**

Figure 5. Overlap of identified unique peptides (a−c) and proteins (d−f) using the five reference database construction method from *M. honghuensis* Liu et Gu, 2011 (a,d), *M. wulii* Landsberg & Lom, 1991 (b,e) and *T. kitauei* Egusa & Nakajima, 1981 (c,f) nematocyst samples.
Table 3. BUSCO Alignment for Assembly Completeness between CCPRD and Alternative Databases: (1) Six-Frame Translation of the Transcriptomes, (2) Six-Frame Translation of the Transcriptome Plus Genomes, (3) Contaminant Database by Adding Host and Bacteria Sequences to the CCPRD, (4) Adding Back Sequences Removed by the Decontamination Process to the CCPRD of M. honghuensis, M. wulii, and T. kitauei

| species     | database name                  | total BUSCO groups searched | total complete (%) | single-copy BUSCOs (%) | duplicated BUSCOs (%) | fragmented BUSCOs (%) | missing BUSCOs (%) |
|-------------|--------------------------------|----------------------------|--------------------|------------------------|-----------------------|----------------------|-------------------|
| M. honghuensis | H_CCPRD             | 978                        | 53.1               | 35.6                   | 17.5                  | 9.6                  | 37.3              |
|             | H_trans_6_frame        | 978                        | 44.7               | 38.0                   | 6.7                   | 9.3                  | 46.0              |
|             | H_genome+transcriptome_6_frame | 978                        | 45.7               | 38.1                   | 7.6                   | 11.9                 | 42.4              |
|             | H_CCPRD_contam         | 978                        | 99.9               | 55.1                   | 44.8                  | 0.1                  | 0.0               |
|             | H_CCPRD_remove         | 978                        | 55.6               | 37.2                   | 18.4                  | 9.9                  | 34.5              |
| M. wulii    | W_CCPRD               | 978                        | 58.2               | 28.1                   | 30.1                  | 8.3                  | 33.5              |
|             | W_trans_6_frame        | 978                        | 51.6               | 36.4                   | 15.2                  | 7.6                  | 40.8              |
|             | W_genome+transcriptome_6_frame | 978                        | 53.4               | 35.9                   | 17.5                  | 9.6                  | 37.0              |
|             | W_CCPRD_contam         | 978                        | 99.9               | 52.0                   | 47.9                  | 0.1                  | 0.0               |
|             | W_CCPRD_remove         | 978                        | 58.4               | 28.2                   | 30.2                  | 9.1                  | 32.5              |
| T. kitauei  | T_CCPRD               | 978                        | 54.9               | 24.2                   | 30.7                  | 9.5                  | 35.6              |
|             | T_trans_6_frame        | 978                        | 48.9               | 27.7                   | 21.2                  | 10.9                 | 40.2              |
|             | T_genome+transcriptome_6_frame | 978                        | 51.5               | 29.2                   | 22.3                  | 11.5                 | 37.0              |
|             | T_CCPRD_contam         | 978                        | 99.9               | 55.1                   | 44.8                  | 0.1                  | 0.0               |
|             | T_CCPRD_remove         | 978                        | 54.9               | 24.2                   | 30.7                  | 9.7                  | 35.4              |

3.5% more proteins compared to the databases trans_6_frame, genome+transcriptome_6_frame, and CCPRD_remove, respectively. The database CCPRD_contam identified the most proteins in three species, while the database trans_6_frame identified the least amount of proteins in M. honghuensis. The database genome+transcriptome_6_frame identified the least amount of proteins in M. wulii and T. kitauei. Adding contaminant data enabled the CCPRD_contam database to identify 33.1, 8.2, and 20.8% more proteins in M. honghuensis, M. wulii, and T. kitauei. As was the case with the peptides, the trans_6_frame database identified more proteins than genome+transcriptome_6_frame in M. wulii (7.4%) and T. kitauei (5.9%), whereas the M. honghuensis trans_6_frame identified fewer proteins (10.2%) than the database genome+transcriptome_6_frame.

The composition details from identification results of CCPRD, CCPRD_contam, and CCPRD_remove are summarized in Table 2.

It was found that, for both the peptide and protein results, CCPRD_contam and CCPRD_remove identified fewer parasite sequences than the CCPRD. The database CCPRD_remove only contained a few identifications from sequences removed during the decontamination process, for example, 102/3126 and 50/1054 for the peptide and protein result for M. honghuensis. The database CCPRD_contam, however, contained a relatively high ratio of off-target hits, for example, 954/3953 and 424/1432 for the peptide and protein result for M. honghuensis. It was also noted that when the database CCPRD_contam was used to study M. honghuensis and M. wulii, there were more bacterial sequences than host sequences. While for T. kitauei, when the function CCPRD_contam was used, there were more host sequences than bacterial sequences.

Overlap of MS Results Using CCPRD and Alternative Databases. Protein results were clustered into ortholog groups (OGs), then compared and illustrated with the Venn diagram (Figure 5).

The core peptides shared by all five databases constituted only 20.96–23.8% of CCPRD OGs. For proteins, this number is 52.3–60.2%. The common peptide OGs between the CCPRD and the database genome+transcriptome_6_frame constituted 24.4, 24.0, and 25.4% of each of the CCPRD OGs for M. honghuensis, M. wulii, and T. kitauei, respectively. For protein OGs shared by the CCPRD and the database genome+transcriptome_6_frame, these numbers were 70.4, 68.9, and 57.7%. Similarly, the shared peptide OGs between the CCPRD and the database trans_6_frame were 27.9, 27.6, and 30.1% for each of the CCPRD OGs (Figure 5a–c). For protein OGs shared between the CCPRD and the database trans_6_frame, these numbers were 91.9, 90.8, and 89.7% (Figure 5d–f). There were 1626, 920, 773 peptides and 136, 49, 52 proteins identified only in the genome+transcriptome_all_6_frame database, but not in the mRNA-only database for M. honghuensis, M. wulii, and T. kitauei, respectively. Besides, the database CCPRD_contam had the most specific OGs in all the species, which comprised 72.5–75.5% of their total OGs for peptides and 20.3–21.8% for proteins.

Completeness Assessment of CCPRD and Alternative Databases. BUSCO attempts to quantitatively assess the completeness of gene set in terms of matching to expected single-copy orthologs. The resulting matches are classified as: (1) “complete” if their lengths are within the expectation of the BUSCO profile match lengths, (2) “duplicated” if these are found more than once, (3) “single-copy” if these are found only once, (4) “fragmented” if the matches that are only partially recovered, (5) “missing” if there are no matches that pass the tests of orthology. Results show that the CCPRD_contam of M. honghuensis, M. wulii, and T. kitauei had the highest level of BUSCO (99.9%) among all of the databases, followed by CCPRD_remove (55.6, 58.4, 54.9%), CCPRD_contam (53.1, 58.2, 54.9%), genome+transcriptome_6_frame (45.7, 53.4, 51.5%), and trans_6_frame (44.7, 51.6, 48.9%) (Table 3).

Apart from the almost complete CCPRD_contam, all the other four databases had relatively high missing BUSCOs: 34.5–46.0, 32.5–40.8, and 35.4–40.2% for M. honghuensis, M. wulii, and T. kitauei. CCPRD_contam of all three species had...
99.9% BUSCO, which are the highest level of completeness among all the databases.

**Overall Peptide Validation by the Linear Fitting of the Actual Peptide Retention Time and Theoretical Hydrophobicity Index.** For all three species, the validity of the peptide identifications was successfully proved, with a strong correlation ($R^2 = 0.891−0.934$) between the actual peptide retention time and theoretical hydrophobicity index (Figure 6).
It was also found that, for all three species, the CCPRD linear fitting had the highest $R^2$ compared with alternative databases.

**Database-Specific Peptide Results Validation.** The linear fitting results of database-specific peptides were identical to the ones reported in above overall peptide validation. For all three species, a strong correlation ($R^2 = 0.908$–$0.928$) between the actual peptide retention time and theoretical hydrophobicity index was observed (Figure S1). The CCPRD linear fitting also had the highest $R^2$-squared compared with alternative databases. The comparison results of the protein reference databases in the number of identified nonredundant peptides showed that, for all three species, the CCPRD still enabled the identification of more nonredundant peptides compared with the databases trans_6_frame, genome+transcriptome_6_frame, and CCPRD_remove (Figure S2). For all three species, the CCPRD_contam identified the most nonredundant peptides among the five databases, while the database genome+transcriptome_6_frame identified the least nonredundant peptides.

**Validation of CCPRD Benchmark in Baker’s Yeast Saccharomyces cerevisiae.** For *S. cerevisiae*, the CCPRD enabled the identification of more peptides than the databases Swiss-Prot, trans_6_frame, genome+transcriptome_6_frame, and CCPRD_remove (Figure 7a).

While for protein results, Swiss-Prot database identified the most among five databases (Figure 7b). The number of identified proteins is only slightly higher than that of the CCPRD (1882 vs 1862). For both peptide and protein results, Genome+transcriptome_6_frame database ranked the third in the identification numbers, followed by CCPRD_contam and trans_6_frame. For overlap results, the core peptides shared by all five databases constituted only 8.43–12.90% of CCPRD OGs (Figure 7c). For proteins, this number is 67.75–87.99% (Figure 7d).

The validity of peptide identifications was proved by the strong correlation ($R^2 = 0.952$–$0.956$) between the actual peptide retention time and theoretical hydrophobicity index (Figure S3). The UniProtKB/Swiss-Prot linear fitting had slightly higher $R^2$-squared than the CCPRD.

**DISCUSSION**

By making the maximum use of genomic and transcriptomic data, we propose a new framework CCPRD for developing a comprehensive, contamination-free, and artifact-free reference protein database for proteomic analyses. The CCPRD has multiple uses for the study of a wide range of species, for which omic resources are suboptimal, involve heavy contamination, or contain complex genomes. This approach may help to develop an optimal reference database for the whole proteome analysis, as we have shown in the case of the myxozoan nematocyst. Our CCPRD analysis of nematocysts illustrates that this approach does work in nonstandard organisms, as well as that the CCPRD can be used with both user-produced or published omic data sets.

One of the most important features of the CCPRD is to filter the contamination in omic data (genomes and transcriptomes). Several efforts had been made to remove contamination using in silico hybridization. For instance, a study compared the genome/transcriptome with host sequences and used an
extremely strict e-value threshold to remove contaminants.24 Although effective, this consideration is of relatively simple dimensions and the e-value is hard to determine in cross-studies. Another work considered the dynamical contaminant and introduced a cross-match method that blasted the myxozoan sequences against host and close-relative databases, respectively. It also removed the sequences by comparing the e-values.25 However, the calculation of the e-value is search-space dependent and not comparable across different databases,30 thus limiting the application of this approach. In the present study, we applied a conservative reciprocal best blast hit (CRBBH) to transcriptomes and the TAGC plot to genomes for decontamination. Both of these were multidimensional and conservative, helping us to avoid overdecontamination. Our results show that the database CCPRD_contam identified many more peptides and proteins than the CCPRD, which suggests the existence of contamination in the source omic data and proves the necessity of decontamination. For the database CCPRD_remove, adding back sequences that were removed during the decontamination process did not increase the number of identified peptides and proteins, which suggests that our method is not overdecontaminated. The slight decrease in identification numbers in the database CCPRD_remove compared with the CCPRD could be explained in that CCPRD_remove did identify peptides not identified by the CCPRD. However, it lost more identifications than it gained, leading to fewer identifications overall. The previously mentioned results indicate that the contamination process in the CCPRD is effective and practical. Our conservative methods have considered the practical aspect of the proteomic study and concluded that the amount number of contaminant proteins is usually fewer in number, which can be easily marked and removed in downstream analyses.

The CCPRD also maximizes the utility of information while saving the search space as much as possible. Traditional proteomic reference databases usually relied on searching MS/MS spectra against public databases, for example, the NCBI nr database and the Swiss-Prot database.31,32 With the development of next-generation sequencing, gene models or transcriptomes have been applied to reference database development, but mostly in the form of three-frame or six-frame translation.14 The CCPRD chose to combine de novo, homolog-based, and the six-frame translation method to maximize the utility of information, and to minimize the nonreal sequences introduced. Our results showed that the CCPRD outperformed the traditional transcriptome six-frame translation method by having identified 35.2% more peptides and 35.8% more proteins, with a maximum of 44% in database size reduction. The CCPRD outperformed the genome plus transcriptome six-frame translation method by having identified 50.7% more peptides and 43.8% more proteins with a maximum of 84.6% in database size reduction. For the database CCPRD_contam, although adding in the contaminant species entries leads to more peptide and protein identifications than the CCPRD, the number of parasite identifications dropped as a result of adding these extra sequences (Table 2). The decrease in the number of parasite identifications was also observed in the database CCPRD_remove, which was likely due to the same reason that the database CCPRD_contam had this outcome. These results demonstrate that the CCPRD is superior to traditional methods in terms of peptides and proteins identification numbers, and database size. The latter is directly related to the amount of time spent searching and available computing resources.18 Besides, it was found that the traditional transcriptome six-frame translation method outperformed the genome plus transcriptome six-frame translation method only in terms of peptide identification numbers. Considering that only the peptide number was proposed as an objective benchmark in the cross-comparison of reference sets,18 and that the database trans_6_frame had a 72.4–86.9% size-reduction rate compared to the database genome+transcriptome_6_frame, the superiority of the transcriptome six-frame translation method over the genome plus transcriptome six-frame translation method was claimed. This suggests that, in some cases (e.g., under-studied species with poor gene models, or genomes with a high level of contamination), adding genomes to the reference database in the traditional context may not necessarily benefit the proteomic study.

In further investigating the overlap of different reference databases, it was found that the pattern of the CCPRD results was more similar to the database trans_6_frame than to the database genome+transcriptome_6_frame, again supporting the harm of adding six-frame translated genomic data (highly fragmented). Yet, there were peptides and proteins only identified in the genome + mRNA-derived database and not in the mRNA-only database, which suggested that there were some present proteins not being actively transcribed. Besides, the MS results obtained by different methods vary widely: simply changing the reference database construction strategies will lead to a huge alteration of identified proteins and peptides. This addresses the important role of the reference database in a proteomic study, as well as the significance of building an efficient and comprehensive database.

Another attractive aspect of the CCPRD is that it maintains a relatively high level of completeness, as identified by a completeness test using the metazoan ortholog set by BUSCO. It was found that, in all species, the CCPRD exhibited a higher completeness level than those from the databases trans_6_frame and genome+transcriptome_6_frame. Nevertheless, most databases used in the study performed poorly under BUSCO analysis: the highest completeness of CCPRD is only 58.2%. It is admitted that high quality proteomics work requires at least a high-quality de novo transcriptome (better still, high-quality genome) to produce useful and interpretable proteins. However, we assumed this not to be a defect in the data itself and will not affect the validity of our results for two reasons: (1) it was noted that both BUSCO and Core Eukaryotic Genes Mapping Approach (CEGMA)33 performed poorly on Myxozoa. CEGMA only recovered 53.6, 37.5, 73, and 46.8% of core genes for Henneguya salminicola, Myxobolus squamalis, Kudoa iwatai, and T. kitauei, respectively.34 This might be because the fast evolutionary rates25 of myxozoans reduced the ability to detect many common eukaryotic genes, a challenge also faced by other fast-evolving eukaryotic lineages.35 (2) The transcriptome sequencings were deep concerning the genome size of myxozoans and the samples were collected from immature plasmodia that contained spores of different stages. Besides, the database CCPRD_contam found that all three species had a 99.9% BUSCO rating, which is the highest level of completeness among all of the databases. This extremely high completeness level is thought to be an artifact because the database CCPRD_contam is served as a contaminant database that introduces almost complete host sequences. It should be noted that, as expected, adding six-frame translated genomic data can improve the completeness of the database trans_6_frame (from 44.7–51.6 to 45.7–53.4%). However, this...
improvement is insignificant, especially considering the sudden increase in the database size. Thus, we concluded that the general performance of trans_6_frame was better than genome +transcriptome_6_frame, second only to the CCPRD.

Next, further validation of the peptide identifications has been done to address two major concerns: (1) it was noted that in Figure 5, the different databases showed limited overlap in the identified peptide sequences. Could this highly varying results be due to the fact that the vast majority of peptide-to-spectrum matches (PSMs) are random and thus incorrect? (2) Even though CCPRD can improve the number of identifications, does it really improve the overall performance, or just increases the number of spurious false positive identifications, or the number of replicate identifications? An important factor in the context of validating PSMs is the difference between the measured retention time and theoretical hydrophobicity index. We used the SSRCalc model to estimate the hydrophobicity index of peptides based on their sequences and chemical modifications, which can be linearly mapped to the retention time. Both the linear fitting results of overall and database-specific peptides suggest that our PSMs are valid and CCPRD can improve the overall validity (higher $R^2$). We thus suggest that the limited overlap in peptides (Figure 5) is not an error and can be explained by the algorithm of clustering: for blast-based program such as OrthoMCL, short sequences cannot produce large bit scores or low e-values, and thus will have low recall rate. This point will not affect our horizontal comparison of difference databases and the conclusion that CCPRD performs better. Besides, the redundancy removal analyses of database-specific peptide identifications revealed that, even with redundancy removed, the number of CCPRD-specific peptides still outperformed the alternative databases. This suggested that the increasing of identifications in CCPRD is not due to the increasing of replicate identifications.

Last but not least, our benchmark is designed for nonmodel organisms, but how will it perform in model organisms? A gold-standard reference with a well-documented and existing protein database will help us to evaluate the utility of CCPRD. The workflow was tested in S. cerevisiae and the performance of CCPRD was assessed. It was found that CCPRD worked well in S. cerevisiae and had similar performance to UniProtKB/SwissProt, a database with well-curated yeast proteins. Furthermore, the overlap of peptide identifications among different reference databases in S. cerevisiae is also limited, thus supporting the above conclusion that the highly varying results in Figure 5 are due to the algorithm of OrthoMCL.

Despite above results, we do not intend to claim that CCPRD is more accurate. Although CCPRD outperforms existing methods in the peptide/protein identification number and database size, this does not mean that the CCPRD could provide higher accuracy. For instance, if an incomplete smaller database is produced by excluding known existing proteins from CCPRD and sent to the same mass spectral matching, more identifications than CCPRD can be acquired at a specific FDR. This does not imply high accuracy. Quite the contrary, these spectra tend to cause many incorrect identifications. Further research is required to incorporate computational algorithms and empirical test into the CCPRD to focus on the accuracy of current benchmark.

**CONCLUSIONS**

Taken together, we consider the CCPRD to be an efficient analytical framework for comprehensive proteomic reference database construction. We demonstrate that, compared to a traditional reference database, the CCPRD enables users to identify more peptides and proteins with a much smaller database size, as well as to acquire a higher level of completeness and better performance in peptide validation. The draft genome required for the CCPRD is highly fragmented, which is practical and affordable for most projects that involve nonmodel organisms. Thus, being able to rapidly assemble the CCPRD with public- or experimentally derived transcriptomic and genomic data is highly desirable, given the number of proteomic projects worldwide. We believe this approach will improve gene and protein identification in a wide range of species, for which little is known about the sequence of translated proteins, as well as provide informative data for their phylogenomics and metaproteomics. Thus, the CCPRD can be a useful tool for molecular evolution and ecological researchers.

**EXPERIMENTAL PROCEDURES**

**Samples.** The research with fish was performed under the National Institute of Health guidelines for the humane use of laboratory animals. All experimental procedures involving fish were approved by the institution’s animal care and use committee of the Huazhong Agricultural University, China. All efforts were made to minimize the suffering of the animals. Briefly, fish infected with myxozoans were sent to the laboratory and kept in a relaying tank prior to being euthanized with an overdose of MS-222 (Sigma-Aldrich, Co., Ltd., St. Louis, MO., USA). Cysts of M. honghuensis Liu et Gu, 2011 and M. wulii Landsberg & Lom, 1991 were obtained from infected allogynogenetic gibel carp Carassius auratus gibelio. Cysts of Thelohanellus kitauei Egusa & Nakajima, 1981 was collected from infected common carp Cyprinus carpio. Myxospores were then purified from the cysts by homogenization, filtering and sucrose gradient centrifugation. Part of purified myxospores were immediately placed into RNAlater (Sigma) or 95% ethanol, frozen in liquid nitrogen and finally stored at −80°C. Other fresh spores were immediately sent to nematocyst isolation as described in technical papers.

**Illumina Sequencing and Assembly.** For RNA-seq, total RNA was isolated from thawed samples (immature plasmodia that contained spores of different stages) with TRIzol (Invitrogen). The purity and integrity of RNA was assessed using NanoPhotometer spectrophotometer (IMPLEN, CA, USA) and RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA) respectively. RNA isolates with an RNA integrity number above 7.0 were sent for RNA library preparation using NEBNext Ultra RNA kit [New England Biolab (NEB), Ipswich, MA] following manufacturer’s recommendations and sequenced as 2 × 125 PE runs with Illumina HiSeq 2500 for M. honghuensis, T. kitauei; as 2 × 150 PE runs with Illumina HiSeq 4000 for M. honghuensis, M. wulii, T. kitauei. The final data size for M. honghuensis were 5.8 Gb (HiSeq 2500) and 23.51 Gb (HiSeq 4000). The size for M. wulii was 14.62 Gb (HiSeq 2500). The size for T. kitauei were 5.8 Gb (HiSeq 2500) and 24.2 Gb (HiSeq 4000). These were deep transcriptome sequencing that achieved saturated coverage, considering the estimated genome size of M. honghuensis (161.09 Mb, estimated from following draft genome), M. wulii (241.8 Mb, estimated from following draft genome), T. kitauei (150.7 Mb). The raw data were trimmed using the trimmomatic, resulting in the removal of key/adaptor sequences, poly-A tails, and low-quality reads (Phred score ≤ 30). The quality of remaining sequences was verified.
using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Filtered reads were de novo assembled through Trinity4^43 and clustered using CD-HIT4^42 and TGICL.4^43

For genomic data, DNA was extracted from the frozen spores of *M. honghuensis* and *M. wulii* using a TIANamp Genomic DNA kit (TIANGEN Inc., Beijing, China). The 500 bp (*M. honghuensis*) and 450 bp (*M. wulii*) libraries were prepared and sequenced on Illumina HiSeq 2500 (2 x 125 PE) and Illumina HiSeq 2500 (2 x 150 PE) respectively. 19.6 Gb and 18.5 Gb clean sequence data were then generated for *M. honghuensis* and *M. wulii*. Raw data were analyzed using the Illumina RTA1.8 analysis pipelines. An initial assembly was produced with SOAPdenovo with default parameters.4^44

Sequence Decontamination. The contamination in transcriptomes and genomes were removed separately using different strategies. To remove potential host and bacterial contamination in transcriptome data, a customized BLAST-based contamination removal method referred to as a: CRBBH, was used. In the first stage of CRBBH, nonredundant host databases containing all available proteins or nucleotide sequences from the host, as well as nonredundant closely related databases containing all available proteins or nucleotide sequences from species closely related to myxozoans, were constructed (in that order) (Figure 1). For the final host nucleotide and protein databases, where the redundancy was removed using CD-HIT (at a threshold of 95%), it was found that each contained 45,486 and 59,764 entries from nucleotide and protein databases, where the redundancy was conserved. This was done to prevent an accidental over-decontamination that could result in a loss of a large portion of actually expressed genes and proteins.

To remove contamination from genomes, Blobtools v1.0.4^6 was used to construct a BlobDB, which contained the genome sequences, coverage, and species information. To achieve this, the genome assemblies were blasted against the NCBI nucleotide database using the BLAST megablast (-culling_limit 5, -e value 1 × 10^-25, -outfmt 6), as well as the UniRef90 using the diamond BLASTX (-sensitive, -k 20, -c 1, -e value 1 × 10^-10). The results were processed by the Blobtools function "create" to annotate each scaffold. A TAGC was drawn at the rank of the phylum and under the taxrule "bestsum". Using the Blobtools function "view", taxonomically annotated non-cnidarian scaffolds with a bit-score of ≥200 were inspected manually and compared against the NCBI nucleotide database (BLASTN, -e value 1 × 10^-5, -max_target_seqs 20, -outfmt 6). Sequences strongly matched to Chordata and Proteobacteria were excluded.

CCPRD and Control Proteomic Reference Database Construction. The CCPRD was built by integrating information from the genomic and RNA-seq data. For genomes, an initial round of de novo gene prediction was carried out with the addition of mapped RNA-seq reads as intron hints to train and run GeneMark-ET,4^48 which provided species HMM file to Augustus4^49 and SNAP5^0 for training and running. A second round of homology-based gene prediction was conducted by GeneWise5^1 with manually selected close relative annotations as protein homology evidence. Finally, transcripts were mapped to the genome by PASA5^2 to further assembly transcriptomes and the best open reading frames were predicted from the representative transcripts generated by PASA. For RNA-seq data, coding sequences within assembled transcripts were firstly predicted by TransDecoder5^3 and GeneMarkS-T5^4 with default parameters. Next, the contigs and singletons were searched against NCBI nr, UniProtKB/Swiss-Prot, eggNOG and KOG databases using BLASTX. The BLASTX reports were parsed for strand, translation frame, e-value, alignment coordinates and bit score in priority of nr, UniProtKB/Swiss-Prot, eggNOG and KOG databases. A maximum of 20 (high-scoring pairs) HSPs were allowed for each hit and only transcripts that showed conserved alignments (e-value ≤ 1 × 10^-5) were retained. The frame shifts and strand reverse of transcripts were detected and corrected according to the BLASTX results. Then, all the qualified transcripts were translated in the hinted frames and corrected coordinates. All the above transcriptome homolog-based steps can be done by a customized Perl script Hercules, which will call the Anvi’o.5^5

Furthermore, those transcripts that were translated neither by de novo nor homolog-based method were translated into amino acid sequences using the Transseq script from the EMBoss framework with the standard codon table. After six-frame translation, we collected the amino acid sequence between stop codons (represented by *) and discarded the sequence contained ambiguous amino acids (represented by X) or its total length was less than 30. Finally, all the proteins resulted from genomic and RNA-seq data were processed by CD-HIT4^42 with a threshold of 100%, to collapse the group into a nonredundant data set. Then, any resulting protein fragments <30 amino acids were discarded, leading to our final CCPRD. The alternative references were generated as follows: we performed the six-frame translation of the transcriptomes (trans_6_frame), and six-frame translation of the transcriptome...
plus genomes (genome+transcriptome_6_frame). Besides, we generated the contaminant database by adding manually selected host and bacteria sequences to the CCPRD (CCPRD_contam). These contaminant proteins were the same from the host protein database, and the bacterial protein database used in above CRBBH decontamination process (Table S1). Besides, we also added back genomic and transcriptomic sequences that removed by the above decontamination process to the CCPRD to create CCPRD_remove. All proteins were also processed by CD-HIT with a threshold of 100%.

**LC and Tandem MS.** Nematocytes were solubilized by adding 400 μL of SDS buffer [4% sodium dodecyl sulfate (SDS), 100 mM Tris-HCl pH 7.6, 0.1 M dithiothreitol] to 1 mL of samples. Vortex and boiling water bath for 5 min. Brief sonication was performed to reduce the viscosity of the lysate. Boiling water bath for 15 min. Then, centrifuge for 40 min at 14,000 g to remove debris by 0.22 μm filters. The filtrate was quantified with the bichinchoninic acid Protein Assay Kit (Bio-Rad, Hercules, CA, USA). 20 μg proteins were mixed with 5× loading buffer boiled for 5 min and then separated on 10% SDS-polyacrylamide gel electrophoresis gel (constant current 14 mA, 90 min) to detect protein purity. Protein bands were visualized with Coomassie Blue R-250 staining. Samples were processed using filter-aided sample preparation. 1800 μL of iodoacetamide buffer (60 mM iodoacetamide in UA), centrifuge at 600 rpm for 1 min, and collect the filtrate and add 40 μL of 0.1% (v/v) HCl, pH 8.5) in a 10 kDa cutoflters. The resulting peptide concentrations were estimated by the UV light spectral density at OD280.

**Validation of CCPRD Benchmark in Baker’s Yeast (S. cerevisiae).** To compare our method with a gold-standard reference, the CCPRD benchmark was further validated in the *S. cerevisiae* model system. Briefly, the public data of *S. cerevisiae* S288C were used, which included Ensembl annotated version of S288c genome R64-1-1, GenBank GCA_000146045.2, and transcriptome shotgun assembly, GenBank GFJR01000000. We applied the above-mentioned CCPRD benchmark on *S. cerevisiae* data and constructed the CCPRD and control proteomic reference database for comparison. The alternative references included the six-frame translation of the transcriptomes (trans_6_frame), six-frame translation of the transcriptome plus genomes (genome+transcriptome_6_frame), the CCPRD + contaminant database with manually selected bacteria sequences (CCPRD_contam), and the UniProtKB/Swiss-Prot database (database release version of May 2020 containing 6,721 yeast protein sequences, www.uniprot.org). The original CCPRD_remove database was excluded in *S. cerevisiae* analyses because almost no contaminants were detected in genomic and transcriptomic data. Then, *S. cerevisiae* spectra raw file was downloaded from the PRIDE database (PXD009214) and searched against the CCPRD and alternative databases using MaxQuant. To evaluate the performance of the CCPRD in *S. cerevisiae*, we compared the results of the CCPRD and the other four alternative reference databases in six aspects. We first compared the number of identified unique peptides and proteins in different reference databases. Second, using OrthoMCL, we compared the overlap of identified unique peptides and proteins between them. Third, BUSCO was used to assess the completeness of the CCPRD and alternative reference databases by mapping against genes that are highly conserved in eukaryotes. Then, the overall peptide identities was validated by the linear fitting of the actual peptide retention time and theoretical hydrophobicity index, which was calculated by SSMCalc model. Finally, we conducted both SSMCalc and redundancy removal analyses (CD-HIT, -c 1 -n 5 -M 16000 -d 0) on the peptides that were identified only when matching against the CCPRD, versus those proteins identified when matching other databases, to check whether the CCPRD actually improved the overall result (rather than increasing the number of spurious false positive identifications, or increasing the number of replicate identifications).

**Database Evaluation.** To evaluate the performance of the CCPRD, we compared the CCPRD and the other four alternative reference databases in six aspects. We first compared the number of identified unique peptides and proteins in different reference databases. Second, using OrthoMCL, we compared the overlap of identified unique peptides and proteins between them. Third, BUSCO was used to assess the completeness of the CCPRD and alternative reference databases by mapping against genes that are highly conserved in eukaryotes. Then, the overall peptide identities was validated by the linear fitting of the actual peptide retention time and theoretical hydrophobicity index, which was calculated by SSMCalc model. Finally, we conducted both SSMCalc and redundancy removal analyses (CD-HIT, -c 1 -n 5 -M 16000 -d 0) on the peptides that were identified only when matching against the CCPRD, versus those proteins identified when matching other databases, to check whether the CCPRD actually improved the overall result (rather than increasing the number of spurious false positive identifications, or increasing the number of replicate identifications).
unique peptides and proteins, the overlap of identified unique peptides and proteins, and the linear fitting of the actual peptide retention time and theoretical hydrophobicity index. The detail procedures are provided as an example file of CCPRD and can be found in the Supporting Information and https://github.com/qingxiangguo/CCPRD.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c01278.

Information of databases used by CRBBH for transcriptome decontamination; linear fitting of measured retention time and theoretical hydrophobicity index of database-specific peptide results of M. honghuensis, M. wulii, and T. kitauei; comparison of the protein reference databases in the number of identified nonredundant peptides for M. honghuensis, M. wulii, and T. kitauei; linear fitting of measured retention time and theoretical hydrophobicity index of baker’s yeast S. cerevisiae peptide results README file of CCPRD (PDF)

**AUTHOR INFORMATION**

Corresponding Author

Zemao Gu — Department of Aquatic Animal Medicine, College of Fisheries, Huazhong Agricultural University, Wuhan, Hubei Province 430070, PR China; Hubei Engineering Technology Research Center for Aquatic Animal Diseases Control and Prevention, Wuhan 430070, PR China; orcid.org/0000-0002-1278-4610; Email: guzemao@mail.hzau.edu.cn

Authors

Qingxiang Guo — Department of Aquatic Animal Medicine, College of Fisheries, Huazhong Agricultural University, Wuhan, Hubei Province 430070, PR China; Hubei Engineering Technology Research Center for Aquatic Animal Diseases Control and Prevention, Wuhan 430070, PR China

Dan Li — Department of Aquatic Animal Medicine, College of Fisheries, Huazhong Agricultural University, Wuhan, Hubei Province 430070, PR China; Hubei Engineering Technology Research Center for Aquatic Animal Diseases Control and Prevention, Wuhan 430070, PR China

Yanhua Zhai — Department of Aquatic Animal Medicine, College of Fisheries, Huazhong Agricultural University, Wuhan, Hubei Province 430070, PR China; Hubei Engineering Technology Research Center for Aquatic Animal Diseases Control and Prevention, Wuhan 430070, PR China

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acsomega.0c01278

Author Contributions

Conceptualization, Q.G. and Z.G.; Methodology, Q.G., Y.Z. and Z.G.; Investigation, Q.G. and D.L.; Resources, Y.Z. and Z.G.; Writing—Original Draft, Q.G. and Z.G.; Writing—review & editing, Q.G., D.L., Y.Z. and Z.G.; Visualization, Q.G., D.L. and Z.G.; Funding Acquisition, Z.G.; Supervision, Z.G.

Notes

The authors declare no competing financial interest. The scripts and README file for CCPRD pipeline are deposited at GitHub (https://github.com/qingxiangguo/CCPRD). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD018851.

**ACKNOWLEDGMENTS**

We acknowledge funding from the Nature Science Foundation of China (grant nos. 31572233 and 31502209), the China Agriculture Research System (grant no. CARS-46), the Hubei Agricultural Sciences and Technology Innovation Center (grant no. 2019-620-000-001-33), and the Featring Talents Cultivation Project (grant no. 4611300108).

**REFERENCES**

(1) Boucher, M. J.; Ghosh, S.; Zhang, L.; Lal, A.; Jang, S. W.; Ju, A.; Zhang, S.; Wang, X.; Ralph, S. A.; Zou, J.; Elias, J. E.; Yeh, E. Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites. *PLoS Biol.* 2018, 16, No. e2005895.

(2) Nyagwange, J.; Tijhaar, E.; Ternette, N.; Mobegi, F.; Tretina, K.; Silva, J. C.; Pelle, R.; Nene, V. Characterization of the Thelera parva sporozoite proteome. *Int. J. Parasitol.* 2018, 48, 265–273.

(3) Tarun, A. S.; Peng, X.; Dumpit, R. F.; Ogata, Y.; Silva-Rivera, H.; Camargo, N.; Daly, T. M.; Bergman, L. W.; Kappe, S. H. I. A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc. Natl. Acad. Sci.* 2008, 105, 305–310.

(4) Wang, T.; Ma, G.; Ang, C.-S.; Korhonen, P. K.; Xu, R.; Nie, S.; Koehler, A. V.; Simpson, R. J.; Greening, D. W.; Reid, G. E.; Williamson, N. A.; Gasser, R. B. Somatic proteome of Haemonchus contortus. *Int. J. Parasitol.* 2019, 49, 311–320.

(5) Swearingen, K. E.; Lindner, S. E. Plasmid parasites viewed through proteomics. *Trends Parasitol.* 2018, 34, 945–960.

(6) Cowell, A. N.; Istvan, E. S.; Lukens, A. K.; Gomez-Lorenzo, M. G.; Vanherschot, M.; Sakata-Kato, T.; Flannery, E. L.; Magistrado, P.; Owen, E.; Abraham, M.; LaMonte, G.; Painter, H. J.; Williams, J. R.; Franco, V.; Linares, M.; Arriaga, I.; Bopp, S.; Corey, V. C.; Gnägid, N. F.; Coburn-Flynn, O.; Reimer, C.; Gupta, P.; Murithi, J. M.; Moura, P. A.; Fuchs, O.; Sasaki, E.; Kim, S. W.; Teng, C. H.; Wang, L. T.; Akidil, A.; Adjalley, S.; Willis, P. A.; Siegel, D.; Tanaseichuk, O.; Zhong, Y.; Zhou, Y.; Linnás, M.; Ollitte, S.; Gamo, F.-J.; Lee, M. C. S.; Goldberg, D. E.; Fidock, D. A.; Wirth, D. F.; Winzeler, E. A. Mapping the malaria parasite druggable genome by using in vitro evolution and Chemo- genomics. *Science 2018*, 359, 191–199.

(7) Eichinger, L.; Pachebat, J. A.; Glöckner, G.; Rajandream, M.-A.; Sengupta, R.; Berriman, M.; Song, C.; Olsen, R.; Szafirski, K.; Xu, Q.; Tunggal, B.; Kummerfeld, S.; Madera, M.; Konfortov, B. A.; Rivero, F.; Bankier, A. T.; Lehmann, R.; Hamlin, N.; Davies, R.; Gaudet, P.; Fey, P.; Pilcher, K.; Chen, G.; Saunders, D.; Sodergren, E.; Davis, P.; Kerhornou, A.; Nie, X.; Hall, N.; Anjard, C.; Hempfling, L.; Bason, N.; Farbrother, P.; Desany, B.; Just, E.; Morio, T.; Rost, R.; Churcher, C.; Cooper, J.; Haydock, S.; van Driessche, N.; Cronin, A.; Goodhead, I.; Munzy, D.; Mourier, T.; Pain, A.; Lu, M.; Harper, D.; Lindsay, S.; Hauser, H.; James, K.; Quiles, M.; Madan Babu, M.; Saito, T.; Buchrieser, C.; Wardropes, A.; Felder, M.; Thangavelu, M.; Johnson, D.; Knights, A.; Lousegeld, H.; Mungall, K.; Oliver, K.; Price, C.; Quail, M. A.; Unshishita, H.; Hernandez, J.; Rabinowitsch, E.; Steffen, D.; Sanders, M.; Ma, J.; Kohara, Y.; Sharp, S.; Simmonds, M.; Spiegel, S.; Tivey, A.; Sugano, S.; White, B.; Walker, D.; Woodward, J.; Winckler, T.; Tanaka, Y.; Shaluskys, G.; Schleicher, M.; Weinstock, G.; Rosenthal, A.; Cox, E. C.; Chisholm, R. L.; Gibbs, R.; Loomis, W.; Plater, M.; Kay, R. R.; Williams, J.; Dear, P. H.; Noegele, A. A.; Barrell, B.; Kuspa, A. The genome of the social amoeba Dictyostelium discoideum. *Nature 2005*, 435, 43.
human serum samples. Depletion strategies for improved resolution in proteomic analysis of Plasmodium falciparum. Nature 2002, 419, 498.

Björnhall, K.; Milioti, T.; Davidson, P. Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. Proteomics 2005, 5, 307−317.

Dowd, W. W.; Wood, C. M.; Rajmuru, M.; Walsh, P. J.; Kültz, D. Natural feeding influences protein expression in the dogfish shark rectal gland: a proteomic analysis. Comp. Biochem. Physiol., Part D: Genomics Proteomics 2008, 3, 118−127.

Kim, S.-J.; Kweon, O.; Cerniglia, C. E. Proteomic applications to elucidate bacterial aromatic hydrocarbon metabolic pathways. Curr. Opin. Microbiol. 2009, 12, 301−309.

Laakmann, S.; Gerds, G.; Erler, R.; Knebelbelsberger, T.; Martinez Ariizu, P.; Raupach, M. J. Comparison of molecular species identification for North Sea calanoid copepods (crustacea) using proteome fingerprints and DNA sequences. Mol. Ecol. Resour. 2013, 13, 862−876.

Hittinger, C. T.; Johnston, M.; Tossberg, J. T.; Rokas, A. Leveraging skewed transcript abundance by ma-seq to increase the genomic depth of the tree of life. Proc. Natl. Acad. Sci. 2010, 107, 1476−1481.

Kellor, D. S.; Provost, E.; Chaerkady, R.; Muthusamy, B.; Manda, S. S.; Subbannayya, T.; Selvan, L. D. N.; Wang, C.-H.; Datta, K. K.; Woo, S.; Diwedi, S. B.; Renuse, S.; Getnet, D.; Huang, T.-C.; Kim, M.-S.; Pinto, S. M.; Mitchell, C. J.; Madugundu, A. K.; Kumar, P.; Sharma, J.; Advari, J.; Dey, G.; Balakrishnan, L.; Syed, N.; Nanjappa, V.; Subbannayya, Y.; Goel, R.; Keshava Prasad, T. S.; Bafna, V.; Sirdeshmukh, R.; Gowda, H.; Wang, C.; Leach, S. D.; Pandey, A. Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. Mol. Cell. Proteomics 2014, 13, 3184−3198.

Krug, K.; Carpy, A.; Behrends, G.; Matic, K.; Soares, N. C.; Macek, B. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteomic experiments. Mol. Cell. Proteomics 2013, 12, 3420−3430.

Wang, X.; Zhang, B. CustomProDB: An R package to generate customized protein databases from ma-seq data for proteomics search. Bioinformatics 2013, 29, 3235−3237.

Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Desser, S. S.; Siddall, M. E. In silico hybridization enables transcriptomic illumination of the nature and evolution of Myxozoa. BMC Genomics 2015, 16, 840.

Leveraging skewed transcript abundance by rna-seq to increase the genomic depth of the tree of life. Proc. Natl. Acad. Sci. 2010, 107, 1476−1481.

Krug, K.; Carpy, A.; Behrends, G.; Matic, K.; Soares, N. C.; Macek, B. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteomic experiments. Mol. Cell. Proteomics 2013, 12, 3420−3430.

Wang, X.; Zhang, B. CustomProDB: An R package to generate customized protein databases from ma-seq data for proteomics search. Bioinformatics 2013, 29, 3235−3237.

Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Desser, S. S.; Siddall, M. E. In silico hybridization enables transcriptomic illumination of the nature and evolution of Myxozoa. BMC Genomics 2015, 16, 840.

ACS Omega 2020, 5, 15370−15384
software system for fast clustering of large est datasets. *Bioinformatics* 2003, 19, 651–652.
(44) Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 2012, 1, 18.
(45) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinf.* 2009, 10, 421.
(46) Laetsch, D. R.; Blaxter, M. L. BlobTools: interrogation of genome assemblies. *F1000Research* 2017, 6, 1287.
(47) Buchfink, B.; Xie, C.; Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 2015, 12, 59.
(48) Lomsadze, A.; Burns, P. D.; Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 2014, 42, No. e119.
(49) Stanke, M.; Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 2003, 33, W465–W467.
(50) Korf, I. Gene finding in novel genomes. *BMC Bioinf.* 2004, 5, 59.
(51) Birney, E.; Clamp, M.; Durbin, R. GeneWise and Genomewise. *Genome Res.* 2004, 14, 988–995.
(52) Rhind, N.; Chen, Z.; Yassour, M.; Thompson, D. A.; Haas, B. J.; Habib, N.; Wapinski, I.; Roy, S.; Lin, M. F.; Heiman, D. I.; Young, S. K.; Furuya, K.; Guo, Y.; Pidoux, A.; Chen, H. M.; Robbertse, B.; Goldberg, J. M.; Aoki, K.; Bayne, E. H.; Berlin, A. M.; Desjardins, C. A.; Dobbs, E.; Dukaj, L.; Fan, L.; FitzGerald, M. G.; French, C.; Gujja, J.; Hansen, K.; Keifenheim, D.; Levin, J. Z.; Mosher, R. A.; Muller, C. A.; Pfeuffer, J.; Priest, M.; Russ, C.; Smialowska, A.; Swoboda, P.; Sykes, S. M.; Vaughan, M.; Vengrova, S.; Yoder, R.; Zeng, Q.; Allshare, R.; Baulcombe, D.; Birren, B. W.; Brown, W.; Ekwall, K.; Kellis, M.; Leatherwood, J.; Levin, H.; Margalit, H.; Martienssen, R.; Nieduszynski, C. A.; Spatafora, J. W.; Friedman, N.; Dalgaard, J. Z.; Baumann, P.; Niki, H.; Regev, A.; Nusbaum, C. Comparative functional genomics of the fission yeasts. *Science* 2011, 332, 930–936.
(53) Haas, B. J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P. D.; Bowden, J.; Couger, M. B.; Eccles, D.; Li, B.; Lieber, M.; MacManes, M. D.; Ott, M.; Orvis, J.; Pochet, N.; Strozzi, F.; Weeks, N.; Westerman, R.; William, T.; Dewey, C. N.; Henschel, R.; LeDuc, R. D.; Friedman, N.; Regev, A. De novo transcript sequence reconstruction from RNA-Seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* 2013, 8, 1494.
(54) Tang, S.; Lomsadze, A.; Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* 2015, 43, No. e78.
(55) Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Soogin, M. L.; Delmont, T. O. Anvi’o: an advanced analysis and visualization platform for ‘omics data. *PeerJ* 2015, 3, No. e1319.
(56) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 2000, 16, 276–277.
(57) Wierniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* 2009, 6, 359.
(58) Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 2016, 11, 2301–2319.
(59) Li, L.; Stoeckert, C. J.; Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003, 13, 2178–2189.
(60) Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015, 31, 3210–3212.
(61) Sarkizova, S.; Klaege, S.; Le, P. M.; Li, L. W.; Oliveira, G.; Keshliahian, H.; Hartigan, C. R.; Zhang, W.; Braun, D. A.; Ligon, K. L.; Bachireddy, P.; Zervantonakis, I. K.; Rosenbluth, J. M.; Ouspenskaya, T.; Law, T.; Justesen, S.; Stevens, J.; Lane, W. J.; Eisenhaure, T.; Lan Zhang, G.; Clauser, K. R.; Hacohen, N.; Carr, S. A.; Wu, C. J.; Reskin, D. B. A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* 2020, 38, 199–209.
(62) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kandu, D. J.; Inugunti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, B.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Andain, E.; Waler, M.; Jamuczak, A. P.; Teneret, T.; Brazma, A.; Vizcaíno, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* 2019, 47, D442–D450.