Published online 24 April 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 2 **1**
https://doi.org/10.1093/nargab/lqad039

Quasi-prime peptides: identification of the shortest peptide sequences unique to a species

Ioannis Mouratidis [1,2], Candace S.Y. Chan[3,4], Nikol Chantzi[1], Georgios Christos Tsiatsianis[1,5], Martin Hemberg [6,*], Nadav Ahituv [3,4,*] and Ilias Georgakopoulos-Soares [1,*]

[1]Department of Biochemistry and Molecular Biology, Penn State College of Medicine, Hershey, PA, USA, [2]Department of Engineering Science, KU Leuven, Leuven, Belgium, [3]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA, [4]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA, [5]National Technical University of Athens, School of Electrical and Computer Engineering, Athens, Greece and [6]Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, USA

Received December 07, 2022; Revised March 02, 2023; Editorial Decision March 29, 2023; Accepted April 06, 2023

ABSTRACT

Determining the organisms present in a biosample has many important applications in agriculture, wildlife conservation, and healthcare. Here, we develop a universal fingerprint based on the identification of short peptides that are unique to a specific organism. We define quasi-prime peptides as sequences that are found in only one species, and we analyzed proteomes from 21 875 species, from viruses to humans, and annotated the smallest peptide kmer sequences that are unique to a species and absent from all other proteomes. We also perform simulations across all reference proteomes and observe a lower than expected number of peptide kmers across species and taxonomies, indicating an enrichment for nullpeptides, sequences absent from a proteome. For humans, we find that quasi-primes are found in genes enriched for specific gene ontology terms, including proteasome and ATP and GTP catalysis. We also provide a set of quasi-prime peptides for a number of human pathogens and model organisms and further showcase its utility via two case studies for *Mycobacterium tuberculosis* and *Vibrio cholerae*, where we identify quasi-prime peptides in two transmembrane and extracellular proteins with relevance for pathogen detection. Our catalog of quasi-prime peptides provides the smallest unit of information that is specific to a single organism at the protein level, providing a versatile tool for species identification.

INTRODUCTION

Advances in massively parallel DNA sequencing technologies have enabled the generation of reference genomes for thousands of biological organisms, across viral, archaeal, bacterial and eukaryotic species, enabling projects such as the Darwin Tree of Life Project (1), which aims to sequence >70 000 eukaryotic species. Through these large-scale collaborative efforts, the number of available proteomes has also rapidly increased (2). The availability of multiple, diverse organismal proteomes increases our understanding of evolution, including horizontal and vertical gene transfer, selection pressures, taxonomic diversity, and how species evolved new traits. There are also far ranging implications for human health, agriculture, pathogen diagnostics and wildlife conservation. Specifically, the identification of protein sequences that are species-specific can serve as a fingerprint of an organism in a sample, and can enable advances in pathogen detection and improved diagnostic methods.

DNA oligonucleotide and oligopeptide sequences found at genomic and proteomic sequences respectively are often referred to as kmers (substrings of length k of a sequence). Usage of kmers for research applications and methods development has been particularly popular in recent years due to their inherent advantages in overcoming alignment requirements (3–5). A number of kmer-based applications have been developed, which include applications in sequence quality control (6), in metagenomics classification (5), in evolution studies (7) and to perform genome assembly (8) among others.

The number of possible peptide kmers grows exponentially with kmer length, but the likelihood of occurrence is

*To whom correspondence should be addressed. Email: izg5139@psu.edu
Correspondence may also be addressed to Nadav Ahituv. Email: nadav.ahituv@ucsf.edu
Correspondence may also be addressed to Martin Hemberg. Email: mhemberg@bwh.harvard.edu

not equal between different peptide kmers. The frequency of a peptide kmer depends on a number of factors, including the frequency of the constituent amino acids, their biological roles, or in the case of low frequency or absence, the potentially deleterious effects to the cell or organism (9,10). Peptides that are entirely absent from a proteome are referred to as nullpeptides, and an extreme case of nullpeptides are peptide primes, peptide kmers that are absent from all known proteins (11). In our previous work we identified 140 308 851 nullpeptide primes; 36 081 six amino acids long and 140 272 770 seven amino acids long, and showed that nullpeptides are under negative selection (12).

Here, we identify the shortest set of peptide kmers that are found in only one species and are absent in every other assembled reference proteomes, termed quasi-prime peptides (Figure 1). The detection of quasi-prime peptides could be used to determine the presence of a particular species in a biological sample. Thus, these sequences are ideally suited for the development of antibodies and other assays for discriminating species. In addition, identification of species-specific peptides could also inform us about the acquisition of new traits and evolution. As a proof of concept, we characterize human quasi-prime peptides, the genes that harbor them and their locations in the human proteome. We also provide a set of quasi-prime peptides for 21 human pathogens, including viral, bacterial and eukaryotic species, and for model organisms.

MATERIALS AND METHODS

Proteomic datasets

Reference proteomes were downloaded from UniProt (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Reference_Proteomes_2022_03.tar.gz, Release 2022_03, 19-Sep-2022) and included a total of 344 archaea, 8623 bacteria, 2119 eukaryota and 10 789 viral reference proteomes.

Peptide kmer extraction

A peptide kmer K is found in the proteome P if there is at least one location i in proteome P such as P[i:i + k] = K. The peptide kmer K is absent from proteome P if there is no such location, based on the current reference sequencing of that proteome found in the UniProt database. For each proteome in the database kmer extraction was performed for kmer lengths up to and including seven amino acids.

The expected number of occurrences of each dipeptide was estimated as the product of the frequency of the constituent amino acids. The enrichment was calculated as the observed to the expected ratio (Supplementary Figure S2).

Proteome simulations

Simulated proteomes were generated using Ushuffle (13) for every reference proteome controlling for dipeptide content. For each simulated proteome peptide kmer extraction was performed for kmer lengths up to and including seven amino acids and compared against the number of kmers identified in the reference proteomes. For each peptide kmer, the proportion of reference proteomes and the proportion of simulated proteomes, in which it was found, was compared and significance was estimated using binomial tests with Bonferoni corrections, from which Figures 3A and B were generated. For each reference proteome and each simulated proteome, the number of kmers identified was compared, and significance was estimated using binomial tests with Bonferoni corrections, from which Figure 2D was generated.

Quasi-prime peptide identification

Species quasi-prime peptides were defined as peptide sequences found in only one reference proteome, which were absent in every other proteome analyzed. Taxonomic quasi-prime peptides were defined as peptide sequences present in at least one species of a taxonomy and absent from all species outside that taxonomy.

Definition of species and taxonomic quasi-prime peptides

We have represented the proteome as a set of protein sequences $P = \{B_i \mid B_i = b_1\, b_2 \ldots b_k,\ b_m \in L_p\ for\ m = 1, \ldots, k\}$ where alphabet $L_p = \{G, A, L, M, F, W, K, Q, E, S, P, V, I, C, Y, H, R, N, D, T\}$ constitutes the standard amino acid code. We define $C = \{PN_1, PN_2, \ldots, PN_k\}$ the set of the k proteomes analyzed, where $PN_i$is the proteome of the i-th species analyzed. For the purposes of this paper, a protein kmer is defined as a short amino acid sequence $K = k_1k_2 \ldots k_l$ with length $1 \leq l \leq 7$. We chose 7 as the ending point for our analysis, as we identified that at this kmer length that at this kmer length, every proteome exhibits a non-trivial number of kmers, which would be necessary for subsequent statistical analyses. We define $K \in P$ if and only if there is protein sequence $A = \alpha_1\alpha_2 \ldots \alpha_n \in P$ and $1 \leq i \leq n - l$, such as $\alpha_i\alpha_{i+1} \ldots \alpha_{i+l-1} = k_1k_2 \ldots k_l$. We developed an algorithm in Python that performs an exhaustive search across input sequences A, identifying all protein kmers such as $K \in P$.

Peptide kmer primes are kmers KP such as $KP \notin P\ \forall\ proteome\ P$. We use the entirety of proteomes found in UniProt (see Materials and Methods → Proteomic Datasets) as an approximation of the known proteome space.

Expanding on the notion of primes, in this work we define the set of peptide quasiprimes of a proteome P_i as $Q = \{K \mid K \in P_i \wedge K \notin P_j\ \forall\ j \neq i\}$.

Therefore, quasi-prime peptides represent the shortest oligopeptides, which are unique to a species (Figure 1).

We define as $T = \{Ta_1, Ta_2, \ldots, Ta_n\}$ a taxonomy of all known species. Taxonomic quasi-prime peptides of taxonomy T_j are short amino acid sequences K as defined above, such that $\exists proteome\ P \in T_j : K \in P \wedge K \notin Q\ \forall\ Q \in T_\lambda\ for\ \lambda \neq j$.

Therefore, taxonomic quasi-primes represent the oligopeptides which are unique to a taxonomy, that is they appear at least once in a species belonging to that taxonomy and they do not appear in any species belonging to any other taxonomy.

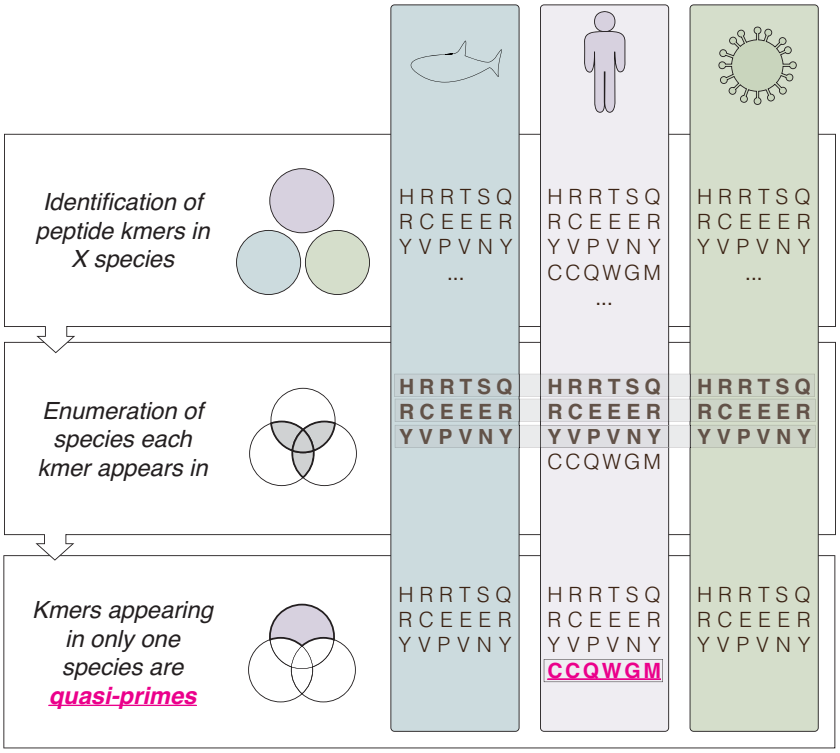

Figure 1. Identification of quasi-prime peptide sequences in individual species. Schematic depicting the detection of six-mer quasi-prime peptide sequences in a species. All kmer peptides of a specific length are identified for each species across all reference proteomes. The kmer peptides that are shared between multiple species are removed and only kmer peptides appearing in a single reference proteome are kept, constituting the set of quasi-prime peptides for that species for that particular length. As an example, we show the hexapeptide CCQWGM that is found in the human reference proteome but is absent from all other species in our database, therefore constituting a human quasi-prime peptide.

GO term analysis

A GO term analysis was performed for genes that contained at least one quasi-prime peptide sequence in the reference human proteome with ShinyGO for GO Biological Process, GO Molecular Process and GO Molecular Function (14). False discovery rate cutoff was set at 0.05, and pathway size minimum of 2 and maximum of 2000 were used.

Gene expression for quasi-prime

Expression levels of every protein-coding gene were measured across 54 tissues using GTeX consortium data RNA consensus tissue gene data (The Human Protein Atlas version 21.1 and Ensembl version 103.38) (15). Proteins were separated into two groups, those containing quasi prime peptides and those not containing quasi prime peptides. Median expression was measured for each group across tissues (Figure 4D) and statistical significance was estimated with Mann–Whitney U tests (P-value < 0.01).

Structure predictions

Structure predictions were retrieved from AlphaFold Protein Structure Database (16) for PPE36 (UniProt accession Q9KMU6) and PRTV (UniProt accession P96811). Visualization was generated using UCSF ChimeraX (17).

RESULTS

Characterization of peptide kmer distribution patterns across taxonomic proteomes

A proteome represents the set of proteins expressed by an organism. We analyzed 21 875 reference proteomes across all domains of life and viruses, processing 25 688 273 765 amino acids. These included reference proteomes for 344 archaea, 8623 bacteria, 2119 eukaryota and 10 789 viruses. We first examined the frequency of each amino acid across all reference proteomes. Even though the most frequent amino acid in the three domains of life and viruses was leucine, we observed distinct patterns of amino acid frequencies between proteomes (Figure 2A). The least frequent amino acid was cysteine, with the exception of eukaryotes, in which tryptophan was the least common (Figure 2A). We performed a principal component analysis (PCA) on the amino acid frequency across reference proteomes and found that the first two principal components (PC1, PC2) explained 57.4% and 10.3% of the variance (Supplementary Figure S1). We also observe that PC2 separates bacteria and eukaryotic proteomes (Supplementary Figure S1). We also compared the frequencies for dipeptides across the domains of life and viruses and found that multiple dipeptides were deviating significantly from the frequencies that would be expected based on their constituent amino acids (Supplementary Figure S2). For instance, the most enriched dipeptides in archaea were Cysteine-Cysteine (CC) and

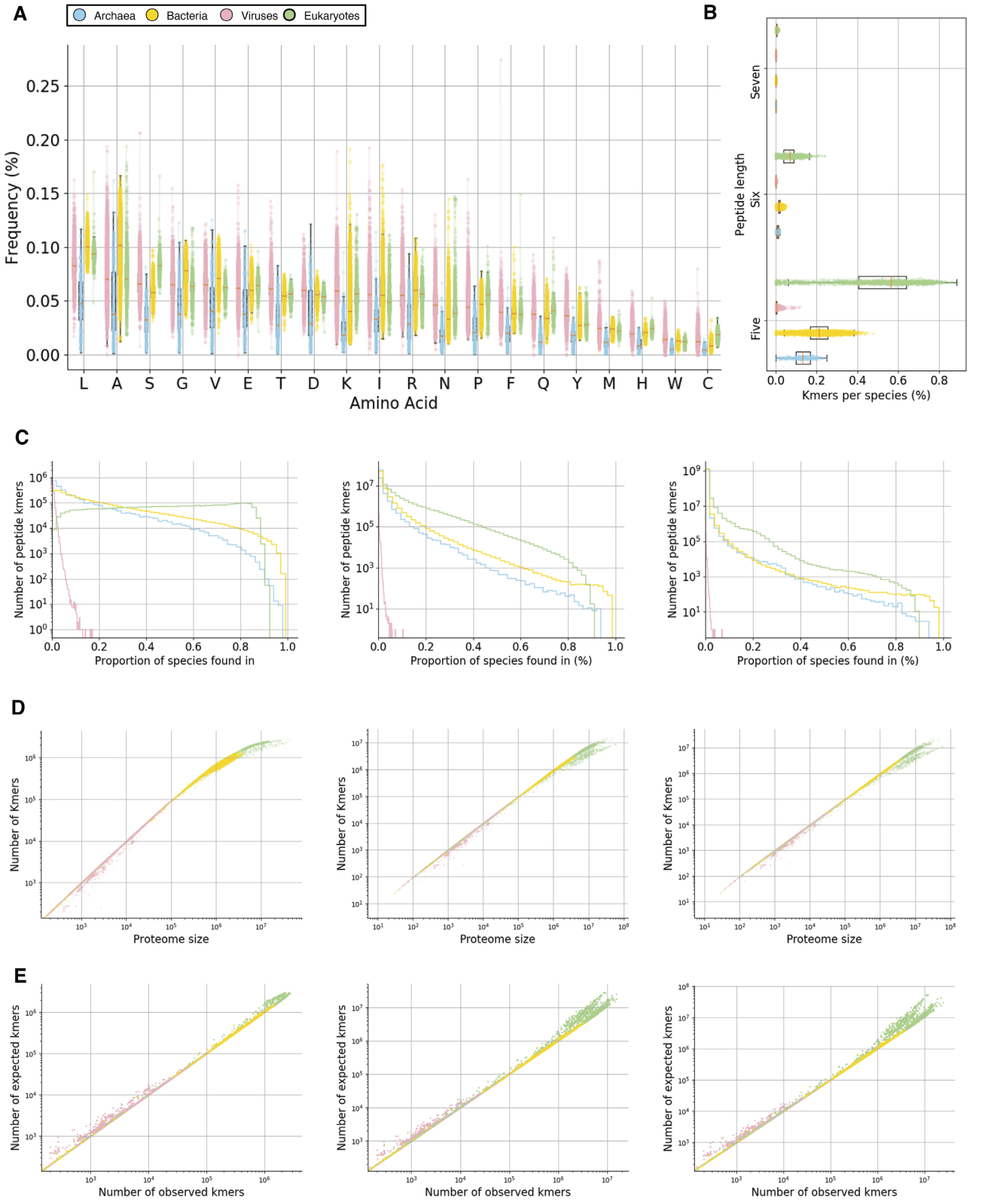

Figure 2. Kmer content as a function of proteome length across organisms. (**A**) Frequency of individual amino acids across taxonomic subdivisions. (**B**) Proportion of kmers found in each species for kmer lengths of five, six and seven amino acids across the taxonomic subdivisions. (**C**) Proportion of species in which each peptide kmer is found in for kmer lengths of five, six and seven amino acids. The majority of peptide kmers are found in a minority of the species studied across taxonomies. (**D**) Number of peptide kmers observed as a function of proteome size for kmer lengths of five, six and seven amino acids. (**E**) Number of expected versus number of observed peptides for each reference proteome for kmer lengths of five, six and seven amino acids. Viral, archaeal, eukaryotic and bacterial proteomes are colored with pink, blue, green and yellow respectively across figure panels.

Cysteine-Proline (CP) with enrichments of 1.65-fold and 1.64-fold respectively. Among all dipeptides, CC was the most enriched dipeptide in both bacteria and viruses. In bacteria CC showed an 1.65-fold enrichment and in viruses it showed an 1.55-fold enrichment (Supplementary Figure S2). For eukaryotes Glutamine-Glutamine (QQ) had the highest enrichment with a 1.41-fold enrichment (Supplementary Figure S2). These results indicate that all three domains of life and viruses have a background biased frequency of amino acids and dipeptides. These findings also indicate a preference for specific dipeptides across multiple domains of life, such as for the CC dipeptide; which likely reflect an evolutionary selection for certain dipeptides in proteins of species across multiple domains of life.

Next, we identified all peptide kmers for lengths up to seven amino acids for all available reference proteomes and examined the distributions of peptide kmer occurrences across the domains of life and viruses (Supplementary Figure S3a). The kmer length of seven amino acids was used as the proportion of kmer space present in each proteome reference decreased to a small number after seven amino acids (Figure 2B, C). We observed that the majority of kmers for five to seven amino acids length are not observed in most species and that the proportion of peptide kmers observed declines with peptide kmer length (Figure 2C). This was particularly the case for kmer lengths of six and seven amino acids, indicating that at longer kmer lengths a larger proportion of the kmer space is absent from a proteome. We also calculated the number of different kmers observed as a function of proteome size for peptide kmer lengths up to seven amino acids (Figure 2D). We observed a strong correlation between proteome size and the number of peptide seven-mers identified across the studied organisms (Pearson $r = 0.922$, P-value < 0.001) and for individual domains, namely archaea (Pearson $r = 0.998$, P-value < 0.001), bacteria (Pearson $r = 0.999$, P-value < 0.001), eukaryotes (Pearson $r = 0.816$, P-value < 0.001) and viruses (Pearson $r = 0.999$, P-value < 0.001).

In our previous work, we examined thirty species and provided evidence for a higher number of nullpeptides than expected by chance (12). However, the aforementioned results do not capture differences between the kmers that were observed in individual organisms, and thus we compared the number of expected versus the number of observed peptide kmers for each species. Across both kmer lengths and organisms we observed fewer kmers than expected, suggesting an enrichment for nullpeptides across species and taxonomic subdivisions (Figure 2E). Specifically, we found 77.62%, 84.60%, 92.10% and 4.91% of archaea, bacteria, eukaryotic and viral species at which there were significantly fewer peptide kmers observed than expected by chance for kmer lengths of seven amino acids (binomial test with Bonferroni corrected P-value, P-value < 0.05), whereas no cases where more kmer peptides than expected by chance were observed. Similar results were also obtained for five and six amino acid kmer lengths. We conclude that there is a higher frequency of nullpeptide sequences than expected by chance across taxonomies, consistent with the notion that these sequences are subject to negative selection.

Enrichment and depletion patterns of individual kmer peptides

Next, we set out to identify enrichment and depletion patterns for peptide kmers at each individual species across all available reference proteomes and across taxonomies. We compared the frequency of each kmer peptide from five to seven amino acids in length at each taxonomic subdivision, to the expected frequency based on permutations of each reference proteome. To that effect, we created a simulated proteome for every single studied species accounting for the dipeptide content of each protein (see Materials and Methods). For every kmer peptide, we calculated the number of reference proteomes it appeared in and contrasted it to the number of simulated proteomes it appears in. For this analysis we only count the number of reference proteomes (or simulations) a kmer peptide appears in and not how many times it appears in total within those species. For a substantial proportion of kmer peptides, we found significant deviations in the number of species they were present in when contrasting reference and simulated proteomes, using binomial tests with Bonferroni correction. For kmer peptides of six amino acids we found 71 486, 121 629, 5 302 237 and 18 306 205 kmer peptides which were differentially enriched in simulations or reference proteomes, representing 0.11%, 0.19%, 8.29% and 28.60% of the kmer space in viruses, archaea, bacteria and eukaryotes respectively (binomial test, Bonferroni correction, P-value <0.05, Figure 3A, B; Table 1). Similar results were found for five and seven amino acid kmers (Table 1). Interestingly, in eukaryotes there were significantly more kmers that were depleted in the reference proteomes relative to the simulated proteomes, whereas in viruses, bacteria and archaea there were more kmers that were enriched in the reference proteomes relative to the simulated proteomes (Figure 3A, B, binomial test, Bonferroni corrected P-values). Archaea and viruses displayed the largest biases, with 99.9% of the peptide kmers with a biased number of occurrences between reference proteomes and simulations, being more frequently observed than expected by chance. These results suggest that for a large proportion of reference proteomes across taxonomic subdivisions the peptide space is being utilized more optimally than expected by chance.

Identification of taxonomic quasi-prime peptides

To identify taxonomic quasi-prime peptides, we annotated quasi-primes that are unique for particular domains of life, meaning peptide sequences that can only be observed in species that are eukaryotic, bacterial, archaeal, or viral, but not in two or more of the aforementioned groups (see Materials and Methods). For kmer lengths of five amino acids, we found that the median number of different kmers observed for viral, archeal, bacterial and eukaryotic proteomes was 4812, 421 305, 676 750 and 1 724 129 respectively (Figure 2C). When combining the kmers observed across species in each taxonomy, the five-mers present in each taxonomy constituted 96.97%, 96.60%, 99.999% and 100% of the five-mer space (Supplementary Figure S3). For peptide kmer lengths of five amino acids, we identified thirteen quasi-prime

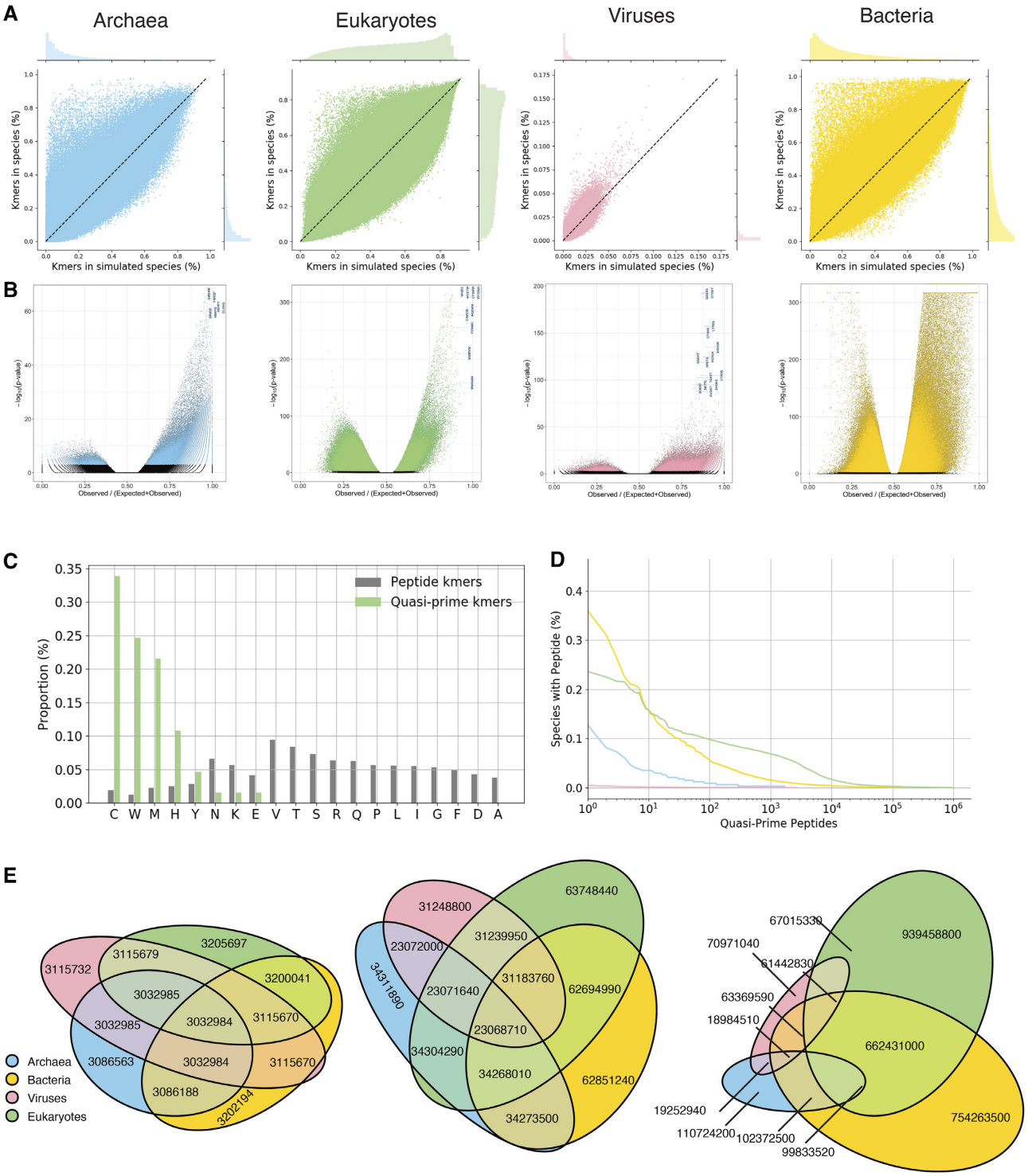

Figure 3. Taxonomic quasi-prime peptide sequences. (**A**) Percentage of reference proteomes and simulated reference proteomes in which each kmer was found for archaea, eukaryotes, viruses and bacteria. (**B**) Volcano plot showing the observed over expected and observed ratio of kmer occurrences in each taxonomy for kmer length of five amino acids. Binomial test with Bonferroni correction was used to estimate statistical significance. (**C**) Proportion of amino acids in observed peptide kmers and eukaryotic quasi-prime peptide kmers for five amino acid kmer length. (**D**) Frequency of taxon-specific quasi-primes in species within that taxon for kmer lengths of six amino acids. (**E**) Venn diagram showing number of shared kmers between taxonomies for kmer lengths of five, six and seven amino acids.

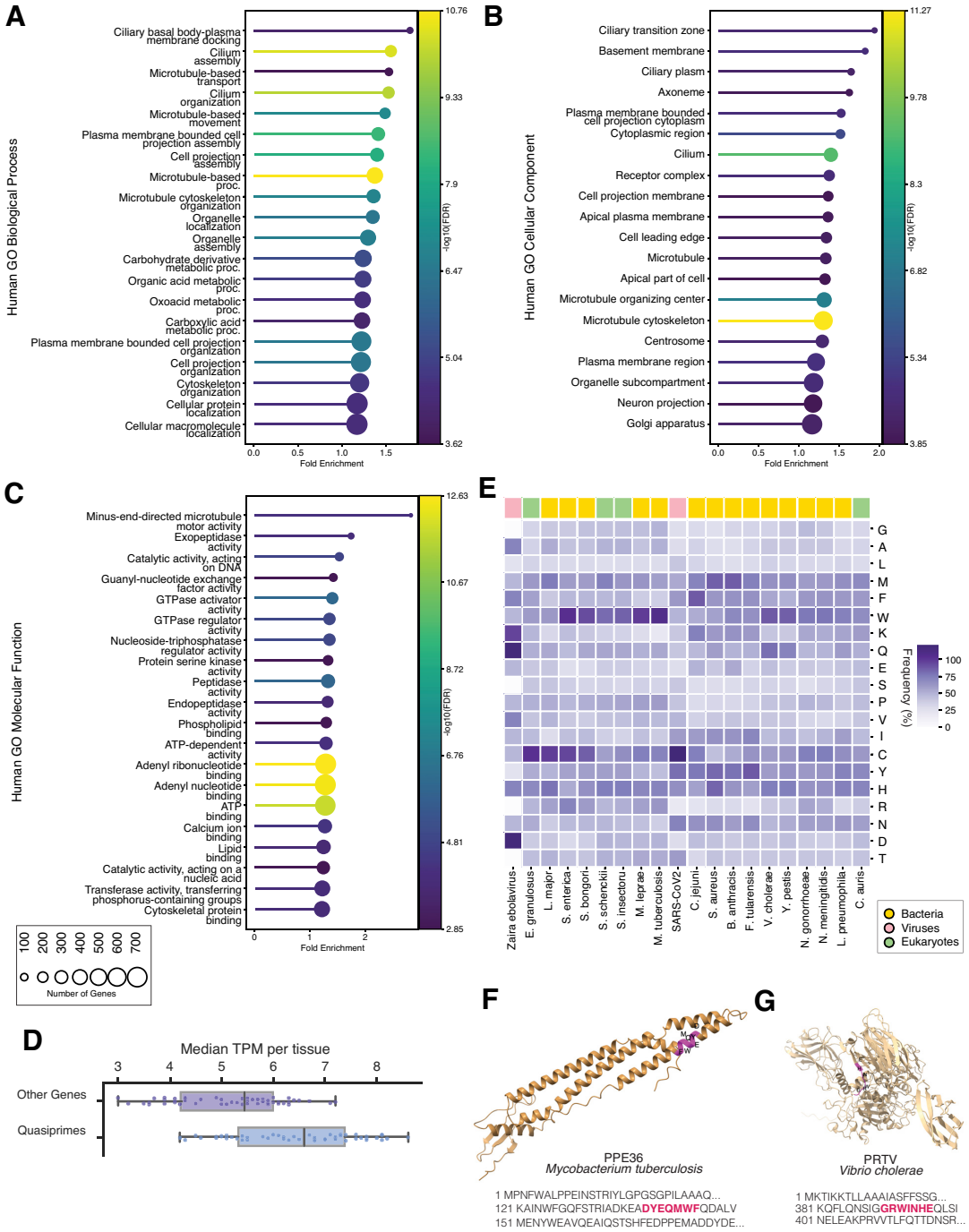

Figure 4. Statistics of quasi-prime sequences in the human proteome. (**A–C**) GO term analysis for human quasi-primes of seven amino acids length. (**A**) Biological processes, (**B**) cellular components and (C) molecular function GO terms are displayed. (**D**) Median expression levels as measured with transcripts per million for genes harboring quasi prime peptides and all other genes in 54 tissues. The results shown are for sevenmer quasi prime peptides. (**E**) Amino acid frequency profile of sevenmer quasi prime peptides across human pathogen species. (**F, G**) AlphaFold structure predictions of PPE36 (F) from *Mycobacterium tuberculosis* and PRTV (G) from *Vibrio cholerae*. The protein sequences along with the ribbon representations of quasi-prime peptides of seven amino acids length (pink).

Table 1. Comparison of kmer frequency in reference proteomes and simulated proteomes in bacteria, archaea, eukaryotes and viruses. Statistical significance was estimated with binomial tests and *P*-values were Bonferroni corrected and significantly biased kmer peptides had adjusted *P*-values < 0.05

Taxonomy	Kmer length	Number of kmers significantly more frequent than expected	Number of kmers significantly less frequent than expected	Total number of kmers with a bias in the number of occurrences
Bacteria	5	387831	731794	1119625
Archaea	5	30056	1110	31166
Viruses	5	33816	4752	38568
Eukaryotes	5	182962	1121963	1304925
Bacteria	6	3289244	2012993	5302237
Archaea	6	121524	105	121629
Viruses	6	71482	4	71486
Eukaryotes	6	3797790	14508415	18306205
Bacteria	7	14974612	4150832	19125444
Archaea	7	607597	321	607918
Viruses	7	241638	8	241646
Eukaryotes	7	15516314	3690345	19206659

peptides in eukaryotes, but no quasi-prime peptides were observed in the other taxonomic subdivisions. The amino acids that were enriched in the thirteen eukaryotic quasi-primes were cysteine, tryptophan, methionine and histidine (Figure 3C).

For kmer lengths of six amino acids, we found that the median number of different kmers observed for viral, archeal, bacterial and eukaryotic proteomes was 4817, 550 801, 1 016 001 and 4 030 763 respectively (Figure 2C). For six-mers present in each taxonomic subdivision we observed that they constituted 48.83%, 53.69%, 98.21% and 99.43% of six-mer space (Supplementary Figure S3). Similarly, for the kmer length of seven amino acids, we found that 5.54%, 8.65%, 58.93% and 73.40% of the kmer space was identified within viral, archeal, bacterial and eukaryotic proteomes respectively (Supplementary Figure S3). We also identified the taxonomic quasi-prime peptide kmers at six amino acids length across the four categories, with 2691 viral, 1715 archaeal, 141 703 bacterial, 954 487 eukaryotic and quasi-prime peptides representing 0.0042%, 0.00268%, 0.22% and 1.49% of the peptide six-mer space. The amino acids that were the most over-represented in taxonomic quasi-prime peptides were cysteine, tryptophan, methionine and histidine across all subdivisions (Supplementary Figure S4).

We also observed large differences in the number of taxonomic quasi prime peptides observed when comparing taxonomies and in the proportion of species harboring each of the taxonomic quasi-primes (Figure 3D, E). For instance, for peptide kmer lengths of six amino acids the viral, archaeal, eukaryotic and bacterial quasi-primes ranged in the proportion of species they were found in (Figure 3D) and the most frequent taxonomic quasi-prime was found in 0.573%, 15.99%, 24.86% and 44.43%, and of the species within the taxonomy respectively, giving us insight about kmers that are very common within a taxonomy but absent from all the others.

Identification of species-specific quasi-prime peptides

We next set out to identify quasi-prime peptides for specific species. To do this, we developed custom algorithms that identify quasi-prime peptides and performed subsequent analyses. The code utilized in this process is available in Github and can be adapted for use with any proteomic dataset (see Materials and Methods).

As a case study, we consider the set of human quasi-prime peptides. We found 107 peptides of length six and 273 741 peptides of length seven that were unique to the human reference proteome. We performed a GO-term analysis for the 7826 genes where these six and seven quasi-peptides appear and found an enrichment for peptidase and ATP and GTP catalysis molecular GO terms and ciliary and microtubule cellular and biological GO terms (Figure 4A–C). In addition, we examined the expression levels of proteins with and without quasi-prime peptides across 54 human tissues. We found that quasi-prime peptides containing proteins show higher expression levels than proteins that do not contain quasi-prime peptides across tissues (Mann–Whitney *U* test, *P*-value < 0.001, Figure 4D). These results were obtained at the transcript level; previously deviations between the relative concentration at the transcript and protein level have been reported (18) and we therefore cannot rule out differences at the proteome level.

We also performed quasi-prime peptide extractions in two other primate species, Bonobo and Gorilla (Supplementary Figure S5), and performed similar GO-term analyses. In Gorilla we found similar GO terms to humans, including ATP and GTP catalysis, cilia and microtubules (Supplementary Figure S6). In Bonobo we observed antigen processing and presentation, cell adhesion, membrane, microtubule and terms associated with ATP and GTP catalysis (Supplementary Figure S6). These results indicate that quasi primes in primates are more frequently observed in proteins with similar processes and function and could be the result of rapid, adaptive evolution at these sequences. We also performed quasi-prime peptide extractions on eight model organisms (Table 2). The number of quasi-prime peptides identified ranged between 19 and 111 for yeast and *C. elegans* respectively at six amino acids kmer length, while for seven amino acids there were several thousand quasi-prime peptides. We also performed GO term analyses for two model organisms namely, yeast and *C. elegans*. In yeast we observe terms associated with cell cycle and plasma membrane among others Supplementary Figure S7a), whereas in *C. elegans*, we observe enrichment for terms associated with detection of chemical stimuli, the plasma membrane and sensory perception (Supplementary Figure S7b). There is evidence supporting strong positive selection for sensory perception genes in *C. elegans* (19), reinforcing the notion that quasi-prime peptides are also associated with rapid, adaptive evolution.

Human pathogen-associated quasi-primes

Quasi primes can serve as an efficient diagnosis tool, allowing the detection of specific pathogens. To showcase this, we next characterized quasi-prime peptides for different types of common or deadly human pathogens, including

Table 2. Quasi-primes in model organisms. The number of quasi-prime peptides for each species across multiple taxonomic subdivisions for kmer peptide length of six amino acids

Species	Number of six-mer quasi-prime peptides	Number of seven-mer quasi-prime peptides
Drosophila melanogaster	28	25463
Arabidopsis thaliana	109	67291
Danio rerio	123	107799
Rattus norvegicus	78	46684
Mus musculus	106	26039
Xenopus tropicalis	104	109883
Saccharomyces cerevisiae	19	19385
Caenorhabditis elegans	111	116664

Table 3. Identification of quasi-primes in human pathogens. The number of quasi-prime peptides for each species across multiple taxonomic subdivisions for kmer peptide length of six and seven amino acids

Species	Taxonomic subdivision	Six-mer quasi-prime peptides	Seven-mer quasi-prime peptides
Salmonella bongori	Bacterium	11	5718
Salmonella enterica	Bacterium	7	2613
Vibrio cholerae	Bacterium	8	7700
Severe acute respiratory syndrome coronavirus 2	Virus	0	131
Neisseria gonorrhoeae	Bacterium	2	1845
Neisseria meningitidis	Bacterium	0	1859
Mycobacterium leprae	Bacterium	2	2137
Necator americanus	Eukaryote	66	58794
Mycobacterium tuberculosis	Bacterium	4	1850
Bacillus anthracis	Bacterium	5	7493
Yersinia pestis	Bacterium	0	7595
Francisella tularensis	Bacterium	4	5221
Staphylococcus aureus	Bacterium	2	5685
Campylobacter jejuni	Bacterium	1	3338
Legionella pneumophila	Bacterium	10	9854
Candida auris	Eukaryote	9	23335
Sporothrix insectoru	Eukaryote	59	43339
Sporothrix schenckii	Eukaryote	30	36957
Leishmania major	Eukaryote	45	27181
Zaire ebolavirus	Virus	0	6
Echinococcus granulosus	Eukaryote	87	67264

6 eukaryotic, 13 prokaryotic and 1 viral species (Table 3). For each of these species, we extracted the set of quasi-prime peptides at six and seven amino acids kmer length (Table 3). We examined the amino acid content of the set of quasi primes in each of these pathogens and observed that across them tryptophan and cysteine were highly enriched (Figure 4E), consistent with our results obtained from taxonomic quasi primes (Supplementary Figures S3 and S4).

We performed two case studies to showcase the existence of quasi-prime peptides in extracellular or transmembrane proteins of human pathogens. For *Mycobacterium tuberculosis*, we identified a quasi-prime peptide in the protein PPE36 (Figure 4F). This protein is a transmembrane protein that is required for heme utilization by *M. tuberculosis* (20), can inhibit M1 polarization of macrophages, is immunogenic against antibodies from sera of pulmonary tuberculosis patients, and reduces the expression of pro-inflammatory cytokines (21). For *Vibrio cholerae,* we found a quasi-prime peptide in the virulence factor PRTV, an extracellular protease that was found to induce cytotoxic effects in human intestinal cell lines and contributes to survival against bacterivorous organisms (22,23) (Figure 4G). These sets of quasi primes provide a dataset for the smallest proteomic unit that is unique to each of these human pathogens.

DISCUSSION

Identification of universal fingerprints could enable the detection of organisms in a sample with increased sensitivity and specificity, with multiple potential applications in agriculture and healthcare. The identification of quasi-prime peptides, which we define and identify here for the first time, enables the derivation of peptide sequences that are unique to a particular species of interest and absent from all other reference proteomes. Therefore, identification of quasi-prime peptides could have direct implications in the development of more reliable and accurate organismal detection methods, such as those using antibodies, mass spectrometry based technologies and biosensors (24). Currently, detection with cell culturing and microscopy based technologies is relatively robust but slow (25,26), whereas immunological-based methods can have limitations in the sensitivity and specificity (27). These limitations often stem from the antigen-antibody reaction and the cross-reactivity of the antibody with other antigens in a sample.

Quasi-prime peptide sequences could enable real-time organismal detection with potentially higher sensitivity and specificity due to their unique peptide sequence content. We generated a quasi-prime profile of multiple model organisms, as well as for a number of human pathogens. In addition, we provide software that can be utilized to generate quasi-primes for any organism of interest. We also find that the number of quasi-prime peptides obtained increases exponentially with peptide length and that the genes that harbor them have associations with particular GO terms. We also define and identify taxonomic quasi-prime peptides, which enable the identification of peptide sequences derived from a set of organisms, which should be absent from organisms which are not part of this taxonomy. Furthermore, we observe that across the examined taxonomies and species, the number of nullpeptides that we observe is larger than expected by chance, supporting our previous observation across 29 proteomes (12), indicating negative selection constraints against certain peptide sequences.

The number of sequenced genomes to date represents only a small fraction of the available genome diversity found in nature. As the number of available reference proteomes will increase over the next few years, additional proteomes could result in the elimination of quasi-prime peptide sequences. However, our computational framework will still be applicable, so the collections presented here can easily be updated. Another limitation specific to the reference proteome analysis for the derivation of quasi-prime peptides is due to alternative splicing and protein isoform formation, which in future work could be accounted for. Also, the addition of missing proteins in future updates of the UniProt database could result in the elimination of a subset of quasi-prime peptides detected. Genomic variants in

populations and sequencing errors could also result in the false identification of quasi-primes. In a practical situation, however, one is likely to rely on hundreds or thousands of peptides and it is unlikely that the potential source of error will cumulatively affect more than a few percent of the quasi-primes for a given species.

DATA AVAILABILITY

The peptide kmer extraction can be performed via the EnumerateNullomers function of the Nullomerator package developed at: https://github.com/Ahituv-lab/Nullomerator.

The code for quasi-prime extraction and subsequent analysis can be found found at: https://github.com/Georgakopoulos-Soares-lab/quasi-prime-peptides and https://doi.org/10.5281/zenodo.7789097.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Author contributions: I.M. and I.G.S. conceived the concept of quasi-prime peptides. M.H., N.A. and I.G.S. supervised the study. I.M., C.S.Y.C., N.C., G.C.T. and I.G.S. wrote the code and performed the analyses. C.S.Y.C. and I.G.S. generated the figures. I.M., C.S.Y.C., G.C.T., M.H., N.A. and I.G.S. wrote the manuscript.

FUNDING

I.G.S. and I.M. were funded by the startup funds of IGS from the Penn State College of Medicine; N.A. and C.S.Y.C. were funded by the National Human Genome Research Institute (NHGRI) [UM1HG011966]; M.H. was funded by startup funds from the Evergrande Center; Helmsley Foundation [2008-04050].
Conflict of interest statement. None declared.

REFERENCES

1. Threlfall,J. and Blaxter,M. (2021) Launching the Tree of Life gateway. *Wellcome Open Res.*, **6**, 125.
2. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
3. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.
4. Alneberg,J., Bjarnason,B.S., de Bruijn,I., Schirmer,M., Quick,J., Ijaz,U.Z., Lahti,L., Loman,N.J., Andersson,A.F. and Quince,C. (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144–1146.
5. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
6. Mapleson,D., Garcia Accinelli,G., Kettleborough,G., Wright,J. and Clavijo,B.J. (2017) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, **33**, 574–576.
7. Bize,A., Midoux,C., Mariadassou,M., Schbath,S., Forterre,P. and Da Cunha,V. (2021) Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. *BMC Genomics*, **22**, 186.
8. Sohn,J.-I. and Nam,J.-W. (2018) The present and future of de novo whole-genome assembly. *Brief. Bioinform.*, **19**, 23–40.
9. Alileche,A. and Hampikian,G. (2017) The effect of Nullomer-derived peptides 9R, 9S1R and 124R on the NCI-60 panel and normal cell lines. *BMC Cancer*, **17**, 533.
10. Tuller,T., Chor,B. and Nelson,N. (2007) Forbidden penta-peptides. *Protein Sci.*, **1s6**, 2251–2259.
11. Hampikian,G. and Andersen,T. (2007) Absent sequences: nullomers and primes. *Pac. Symp. Biocomput.*, 355–366.
12. Georgakopoulos-Soares,I., Yizhar-Barnea,O., Mouratidis,I., Hemberg,M. and Ahituv,N. (2021) Absent from DNA and protein: genomic characterization of nullomers and nullpeptides across functional categories and evolution. *Genome Biol.*, **22**, 245.
13. Jiang,M., Anderson,J., Gillespie,J. and Mayne,M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192.
14. Ge,S.X., Jung,D. and Yao,R. (2020) ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, **36**, 2628–2629.
15. Uhlén,M., Fagerberg,L., Hallström,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,Å., Kampf,C., Sjöstedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
16. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A. *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
17. Pettersen,E.F., Goddard,T.D., Huang,C.C., Meng,E.C., Couch,G.S., Croll,T.I., Morris,J.H. and Ferrin,T.E. (2021) UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.*, **30**, 70–82.
18. Edfors,F., Danielsson,F., Hallström,B.M., Käll,L., Lundberg,E., Pontén,F., Forsström,B. and Uhlén,M. (2016) Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.*, **12**, 883.
19. Ma,F., Lau,C.Y. and Zheng,C. (2021) Large genetic diversity and strong positive selection in F-box and GPCR genes among the wild isolates of Caenorhabditis elegans. *Genome Biol. Evol.*, **13**, evab048.
20. Mitra,A., Ko,Y.-H., Cingolani,G. and Niederweis,M. (2019) Heme and hemoglobin utilization by mycobacterium tuberculosis. *Nat. Commun.*, **10**, 4260.
21. Gong,Z., Han,S., Liang,T., Zhang,H., Sun,Q., Pan,H., Wang,H., Yang,J., Cheng,L., Lv,X. *et al.* (2021) *Mycobacterium tuberculosis* effector PPE36 attenuates host cytokine storm damage via inhibiting macrophage M1 polarization. *J. Cell. Physiol.*, **236**, 7405–7420.
22. Vaitkevicius,K., Rompikuntal,P.K., Lindmark,B., Vaitkevicius,R., Song,T. and Wai,S.N. (2008) The metalloprotease PrtV from *Vibrio cholerae*. *FEBS J.*, **275**, 3167–3177.
23. Vaitkevicius,K., Lindmark,B., Ou,G., Song,T., Toma,C., Iwanaga,M., Zhu,J., Andersson,A., Hammarström,M.-L., Tuck,S. *et al.* (2006) A Vibrio cholerae protease needed for killing of caenorhabditis elegans has a role in protection from natural predator grazing. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 9280–9285.
24. Sfragano,P.S., Moro,G., Polo,F. and Palchetti,I. (2021) The role of peptides in the design of electrochemical biosensors for clinical diagnostics. *Biosensors*, **11**, 246.
25. Iqbal,S.S., Mayo,M.W., Bruno,J.G., Bronk,B.V., Batt,C.A. and Chambers,J.P. (2000) A review of molecular recognition technologies for detection of biological threat agents. *Biosens. Bioelectron.*, **15**, 549–578.
26. Kumar,N., Kumari,M. and Arun,R.K. (2022) Development and implementation of portable biosensors in microfluidic point-of-care devices for pathogen detection. *Miniaturized Biosensing Devices: Fabrication and Applications*. Springer Nature Singapore, Singapore, pp. 99–122.
27. Skottrup,P.D., Nicolaisen,M. and Justesen,A.F. (2008) Towards on-site pathogen detection using antibody-based sensors. *Biosens. Bioelectron.*, **24**, 339–348.