UNIQmin, an alignment-free tool to study viral sequence diversity across taxonomic lineages: a case study of monkeypox virus

Li Chuin Chong[1,2,†] & Asif M. Khan[1,2,*]

[1] Centre for Bioinformatics, School of Data Sciences, Perdana University, Damansara Heights, 50490 Kuala Lumpur, Malaysia

[2] Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Beykoz, 34820 Istanbul, Turkey

* Corresponding author (AMK):- Tel: +90 (212) 523 22 88; Email: asif@perdanauniversity.edu.my, makhan@bezmialem.edu.tr

[†] Present address: Institute for Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, a joint venture between Medical School Hannover (MHH) and Helmholtz Centre for Infection Research (HZI), Hannover, Germany.

Abstract

Sequence changes in viral genomes generate protein sequence diversity that enable viruses to evade the host immune system, hindering the development of effective preventive and therapeutic interventions. Massive proliferation of sequence data provides unprecedented opportunities to study viral adaptation and evolution. Alignment-free approach removes various restrictions, otherwise posed by an alignment-dependent approach for the study of sequence diversity. The publicly available tool, UNIQmin offers an alignment-free approach for the study of viral sequence diversity at any given rank of taxonomy lineage and is big data ready. The tool performs an exhaustive search to determine the minimal set of sequences required to capture the peptidome diversity within a given dataset. This compression is possible through the removal of identical sequences and unique sequences that do not contribute effectively to the peptidome diversity pool. Herein, we describe a detailed four-part protocol utilizing UNIQmin to generate the minimal set for the purpose of viral diversity analyses at any rank of the taxonomy lineage, using the latest global public health threat monkeypox virus (MPX) as a case study. These protocols enable systematic diversity studies across the taxonomic lineage, which are much needed for our future preparedness of a viral epidemic, in particular when data is in abundance and freely available.

Keywords

minimal set, alignment-independent, alignment-free, sequence diversity, proteome, virus, UNIQmin, monkeypox virus

Introduction

Infectious diseases have caused irreversible losses to human lives, contributing greatly to the global disease burden. Amongst pathogenic agents, viruses are abundant and ubiquitous, responsible for the surge in global death toll from infectious diseases. The current coronavirus disease (COVID-19) pandemic is a testimony to this, affecting more than 190 countries/geographical regions, where not just developing countries but even the developed ones with some of the most advanced health systems are crippled and struggling to control the pandemic. Thus far, more than 500 million people have been infected, of which more than six (6) million have died (as of July 2022). Viral diversity, in particular amongst viruses of RNA genetic make-up, poses a significant challenge to the development of diagnostic, therapeutic and prophylactic interventions [1,2]. Viral sequence variability within an infected individual can be an outcome of not just mutation, but also re-assortment and/or recombination, creating a diversity spectrum, termed as viral quasispecies [3], which can consist of one or more variants that are better fit for evolutionary selection. Sequence change, even that of a single amino acid substitution, can lead to immune escape and/or immunopathogenesis in some cases [4].

The advances in genomics and proteomics approaches, coupled with the exponential reduction in the cost of sequencing, have allowed for a massive proliferation of sequence data. This provides unprecedented opportunities to study viral adaptation and evolution. Alignment-dependent approach is typically employed to study viral sequence divergence and conservation [5]. Conserved sequence regions can capture the genotypic diversity of majority or all historically reported variants of a virus, and likely that of future variants [6–9]. Such regions are, however, limited in viral species that are highly variable, namely influenza A virus and human immunodeficiency virus 1 (HIV-1), amongst others, and even more so when applied to the search for universal vaccine targets that capture the diversity of multiple subtypes or subgroups of a highly diverse virus [10,11]. Naturally, aligning a large number of sequences of multiple viral species at the genus or family taxonomic lineage rank can become impractical, corresponding to a decline in the number and length of shared blocks of conserved regions that can anchor the alignment [12]. Separately, multiple sequence alignment requires manual inspection to ascertain reliability, with correction of any misalignment. Further, aligning a large number of sequences can require a significant compute resource [13,14]. Therefore, there is a need for the development of alignment-free or -independent approaches to enable the study of viral sequence diversity at any rank of the taxonomic lineage.

The premise of an alignment-free approach is that it does not rely on the assignment of residue-residue correspondence to quantify sequence similarity. As such, the approach removes various restrictions, otherwise posed by an alignment-dependent approach for the study of sequence diversity [14]. An alignment-free approach can involve performing an exhaustive search to determine the minimal set of sequences required to capture the diversity within a given dataset [12]. The minimal set is herein defined as the smallest possible number of unique sequences required to represent the diversity inherent in the entire repertoire of overlapping *k*-mers encoded by all the unique sequences in a given dataset. The complete set of overlapping *k*-mers can be termed as part of the peptidome of the dataset, and thus the minimal set derived for a specific *k*-mer length is representative of the peptidome diversity relevant to the *k*-mer. Data compression of a given dataset to generate the minimal set is achieved at two levels. First, the redundant reduction (RR) of the dataset or the removal of duplicate sequences, which are common in public databases.

Second, the non-redundant reduction (NRR) of the dataset, which is the removal of unique sequences whose entire repertoire of overlapping *k*-mer(s) can be represented by other unique sequences and thus, rendering them redundant to the collective pool of peptidome sequence diversity relevant to the *k*-mer. The compression can be significant and offer important insights into the effective sequence diversity and evolution of viruses when applied not just at the species level (such as analysis between specific proteins or proteome-wide), but at any rank of viral taxonomy lineage, such as genus, family or even at the highest, super kingdom level (all reported viruses). The study of minimal set has been previously reported for important viruses, such as dengue virus [15,16] and influenza A virus [17].

We have recently developed a novel algorithm for the search of a minimal set [12], which is improved and scalable for massive datasets, compared to the earlier iteration [15,16]. This has been implemented as a tool, UNIQmin to allow for a user-specific search for a minimal set of any *k*-mer at any rank of viral taxonomy lineage. The tool is publicly available via GitHub (https://github.com/ChongLC/MinimalSetofViralPeptidome-UNIQmin) and PyPI (https://pypi.org/project/uniqmin). The utility of the tool was demonstrated for the species *Dengue virus*, genus *Flavivirus*, family *Flaviviridae*, and even at the superkingdom rank, all reported *Viruses*. Herein, we describe a detailed four-part protocol (Figure 1) utilizing UNIQmin to address the issue of analysing viral diversity at any rank of the taxonomy lineage, which is much needed for our future preparedness of a viral epidemic. Monkeypox virus is used as a case study given the recent surge of human cases in countries where the disease was not typically reported.

Methods and Results

Basic Protocol 1: Data Preparation

Viral sequence data, a pre-requisite and key element for the study of diversity, is available in abundance in various public repositories, such as the **N**ational **C**enter for **B**iotechnology **I**nformation (NCBI; https://www.ncbi.nlm.nih.gov/) Entrez Protein (NR) database [18]. Derived or secondary sources, such as the NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/) [19] and **Vi**rus **P**athogen Database and Analysis **R**esource (ViPR; https://www.viprbrc.org/) [20], among others, have also become commonly used resource for viral sequence data. They offer better data integration through other internal and extra sources, and provide internal curation, ease of data visualisation and download via various options for search result customisation. However, they can differ in their depth and breadth of curation, species coverage, and number of sequences. For example, ViPR covers 7,124 species/subgroups (as of January 2022), relative to NCBI Virus and NR, which in contrast cover approximately seven-fold more viral species/subgroups (47,823; as of January 2022); however, for a specific virus species that is covered by all the three databases, the NCBI Virus provides the best user-experience, in general.

Specialist databases are only available for a select few viruses, such as influenza A virus, human immunodeficiency virus (HIV-1) and the virus responsible for the current COVID-19 pandemic, SARS-CoV-2, among others, typically enriched with high curations [21,22]. The **G**lobal **I**nitiative on **S**haring **A**ll **I**nfluenza **D**ata (GISAID; https://www.gisaid.org/), which started with EpiFlu™ database, later expanded to EpiCoV™ and EpiRSV™ and most recently added EpiPox™ boasts 1,651,325, 307,701,631, 24,420 and 482 sequence records (as of July 2022) for influenza viruses, SARS-CoV-2, respiratory syncytial virus (RSV) and monkeypox virus (MPX) respectively, with readily available metadata that facilitates "on-the-fly" analyses. These specialist databases are exemplary in their approach to biological data-warehousing [23,24], such that they have become the preferred repository for community sequence submission, and in the case of GISAID, with numbers that far surpass the traditional primary repositories and even other specialist databases.

The wide availability, low-cost, and portability of next generation sequencing (NGS) offers unprecedented opportunity to the research community for sequencing of viruses of interest, including in real-time at the point/site of sample collection (nanopore technology). It is not uncommon for sequencing projects to output related or identical strains, which contribute to sampling bias [25]. The removal of identical sequences may be necessary for a diversity study, to not only minimize the bias, but also reduce the demand on computational resources. In this protocol, we will showcase a standard workflow for data preparation.

1. Retrieve viral protein sequences from publicly available database of choice.

   Sequence retrieval for a viral species of interest from the primary NCBI NR database is preferably done via the NCBI Taxonomy Browser, by searching for the species taxonomy identifier (txid) or the species name. It is important to ensure that the correct species has been determined from the search, as viruses can have similar names. For example, hepatitis A (HAV), hepatitis B (HBV), and hepatitis C (HCV) viruses, which appear related based on the common names have distinct lineages and thus, scientific names, such as species *Hepatovirus A* of the genus *Hepatovirus*, species *Hepatitis B* virus of the genus *Orthohepadnavirusvirus*, and species *Hepacivirus C* of the genus

Hepacivirus, respectively. Thus, one should check the species lineage and cross-reference with the literature for the specific species of interest.

The species *Monkeypox virus* (MPX) will be used as a case study to demonstrate data retrieval. Navigate the web client to the NCBI Taxonomy Browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) (Figure 2) and search using the respective txid "10244", followed by a download of the sequences in FASTA format.

Sequence data download should be accompanied with metadata download using the "GenPept (full)" format option. The metadata is useful for data processing, especially for filtering irrelevant sequences. The full records of nucleotide sequences should also be downloaded in 'FASTA' format and the metadata in "GenBank (full)" format. They can serve as a reference for comparative analysis.

It should be noted that since no alignment is to be done, both full-length and partial sequences can be included in the dataset for a comprehensive analysis of diversity. In an alignment-dependent study, partial sequences are a common source of spurious alignment, and hence it may be desired that they are filtered out from analyses, particularly when the number of such partial sequences may be prohibitively large, or the protein is of high diversity. Otherwise, in such cases, inspection and correction of misalignment can be a challenge and highly subjective. These concerns are not applicable for an alignment-free approach.

2. [Optional] Extract selected viral protein(s) of interest from the retrieved data

A dataset retrieved from a primary or a secondary data resource(s) for a viral species of interest would comprise sequences of the proteins encoded by the viral genome. It may be in the interest of the user to analyse only one, a group, or all of the proteins encoded. Depending on the architecture of the virus genome, it is possible that a protein record would provide the sequence of only a protein encoded by the genome, as in the case of influenza virus (segmented genome) or provide the sequence of one, more than one, or all the proteins encoded by the genome, as in the case of dengue virus (polyprotein translation of the genome).

A BLAST [26] search can be carried out to facilitate the extraction of the protein of interest. The retrieved dataset is used to construct a searchable database, whilst a reference sequence of the protein of interest can serve as the query for the BLASTp search. The reference sequence can be identified through a search in highly curated protein databases, such as UniProt (https://www.uniprot.org/) [27] or RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) [28].

3. Remove redundant sequences from the dataset to be analysed.

The tool, **C**luster **D**atabase at **H**igh **I**dentity with **T**olerance (CD-HIT; http://weizhong-lab.ucsd.edu/cd-hit/) is ideal for removal of redundant sequences at a given percentage similarity threshold of choice [29]. There are two ways to execute CD-HIT, either locally or using the webserver. One can load the sequence dataset to be analysed in FASTA format to the webserver (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit) and set the parameters. Removal of identical rather than similar sequences is preferred herein, and thus, the sequence identity cut-off should be set to 1, indicating 100% identity. Add desired email address for job checking and click on the ‘Submit’ button. Local client version of CD-HIT is preferred for large datasets.

4. Calculate the percentage of redundant reduction (RR) for the selected dataset analysed.

   Redundant reduction (RR) is defined as the percentage of identical sequences removed from the retrieved dataset and is calculated by use of the equation (Eq.) 1 below:

$$Redundant\ reduction\ (RR) =$$

$$\frac{number\ of\ protein\ sequences\ in\ the\ retrieved\ dataset\ (redundant,\ r\ dataset) - number\ of\ protein\ sequences\ in\ the\ deduplicated\ dataset\ (non-redundant,\ nr\ dataset)}{number\ of\ protein\ sequences\ in\ the\ retrieved\ dataset\ (redundant,\ r\ dataset)}\ X\ 100$$

   For example, a retrieved dataset of 19,423 reported protein sequence of MPX (as of July 2022) was deduplicated using CD-HIT, resulting in a nr dataset of 1,245 sequences, which is a significant RR of ~93.6%.

Basic Protocol 2: Generating a minimal set for any given dataset

This protocol demonstrates the generation of a minimal set for a dataset of interest. As per the definition of the minimal set, the *k*-mer size needs to be defined. Various *k*-mer sizes may be explored; see Chong *et al.* 2021 for the various considerations. Herein, as an example, we will utilise the *k*-mer size of nine (9-mer) for immunological applications, such as studying the viral diversity in the context of the cellular immune response (antigenic diversity). Peptides of different binding length can be recognised by the human leukocyte antigen (HLA) molecules. HLA-I molecules can bind peptides of length, ranging from eight (8) to 15 amino acids (aa), with nine (9) aa being the typical length; HLA class II molecules can bind longer peptides, up to 22 aa, with a binding core of nine (9) aa. The tool UNIQmin will be used for the generation of the minimal set. The algorithm is detailed in Chong *et al.* 2021 and on the GitHub page (https://github.com/ChongLC/MinimalSetofViralPeptidome-UNIQmin/blob/master/README.md).

1. Generate overlapping 9-mers using Step 1 script of UNIQmin, "U1_KmerGenerator.py".

   The non-redundant (nr) file in FASTA format is used as the input to generate a set of overlapping 9-mers. Before executing the "U1_KmerGenerator" script, one may modify the number of CPU-cores to be used, based on the specifications of the in-house machine utilised. Similarly, one may change the *k*-mer size of interest. In this protocol, 14 cores of CPU and *k*-mer of nine are used, while nr file of MPX (MPX_nr.fasta) is used as the input. The command line argument for this step is detailed as below:

   > python U1_KmerGenerator.py MPX_nr.fasta

   One may change the number of CPU-cores to speed up the process for a large dataset, by modifying the code snippet below in the Python script (where variable '**X**', shown in bold and red colour is the placeholder for the number of cores, such as "14" in our case):

```
if __name__ == '__main__':
        n = len(fileA)
        pool = ProcessPoolExecutor(X)
        futures = []
        perCPUSize = math.ceil(n/X)
        for i in range(0,X):
                futures.append(pool.submit(generate_kmers, i*perCPUSize, (i+1)*perCPUSize))
```

   The *k*-mer size can be changed by modifying the code snippet as below (where variable '**N**', shown in bold and red colour is the placeholder, such as "9" in our case):

```
def generate_kmers(fileA, args, file_id, start, end):
        for record in fileA[start:end]:
                nr_sequence = record.seq
                seq_len = len(nr_sequence)
                kmer = N
                count = 0
                temp = []
                for seq in list(range(seq_len-(kmer-1))):
                        count += 1
                        my_kmer = (nr_sequence[seq:seq+kmer])
                        temp.append(str(my_kmer))
                with open(file_id, 'a') as f:
                        f.writelines("%s\n" % kmer for kmer in temp)
```

2. Group the generated overlapping 9-mers from the output file of Step 1 according to their occurrence count, by use of the scripts "U2.1_Singletons.py" and "U2.2_Multitons.py" for single and multi-occurring 9-mer peptides, respectively.

   The output file from Step 1 can comprise of single occurring and multi-occurring 9-mers. The separation between single and multi-occurring peptides is a key attribute of UNIQmin algorithm in striking a balance between speed and accuracy. This separation results in two output files for Step 2, namely "seqSingleList.txt" and "seqmore1List.txt". This step can be executed by use of the command lines below:

   > python U2.1_Singletons.py

   > python U2.2_Multitons.py

3. Identify a pre-qualified minimal set of sequences from the generated output file of Step 2, the one consisting of the single occurring 9-mer peptides ("seqSingleList.txt").

   The "U3.1_PreQualifiedMinSet" script matches all the single occurring 9-mer peptides (in "seqSingleList.txt") to all the input sequences (in MPX_nr.fasta). The matched sequences will be identified and deposited into a new output file, the pre-qualified minimal set file, named as "*file Z*", by use of the command line below:

> *python U3.1_PreQualifiedMinSet.py*

4. Generate a dataset that only contains the remaining input sequences (from MPX_nr.fasta) not included in the pre-qualified minimal set, *file Z* (Step 3).

   By using the "U3.2_UnmatchedSingletons" script, the sequences in the pre-qualified minimal set are removed from the Step 1 input file (MPX_nr.fasta), resulting in an output file with the remaining initial input sequences, which are subsequently to be matched with the multi-occurring 9-mers (Step 7).

   > *python U3.2_UnmatchedSingletons.py*

5. Deduplicate the multi-occurring 9-mer peptides in the output file of Step 2, "seqmore1List.txt" to produce a non-redundant list.

   The output file from Step 2 containing the multi-occurring 9-mer peptides ("seqmore1List.txt") is deduplicated by use of "U4.1_Non-SingletonsDedup" script, executed as below, to produce a non-redundant list of the multi-occurring peptides:

   > *python U4.1_Non-SingletonsDedup.py*

6. Match the non-redundant list of the multi-occurring 9-mers from Step 5 to the pre-filtered minimal set, *file Z* (Step 3) and remove the matching 9-mers.

   This step would result in an output file with the remaining unmatched multi-occurring 9-mers (did not match to sequences in *file Z*).

   > *python U4.2_Multi-OccurringPreMinSet.py*

   > *python U4.3_UnmatchedMulti-Occurring.py*

7. Identify the remaining initial input sequences (of Step 4) that capture the still existing unmatched multi-occurring 9-mers (of Step 6). Such sequences, together with the pre-qualified minimal set, would comprise the final minimal set.

The Step 6, remaining unmatched multi-occurring 9-mers are compared with each of the initial input sequences that are still not part the minimal set (Step 4). All sequences are quantified based on the match and sorted in descending order, from the highest (maximal matching *k*-mer coverage) to the lowest. The sequence with the maximal *k*-mer coverage is deposited into the minimal set, *file Z*. The deposited sequence and the cognate captured 9-mers are then removed from their respective files of this step. This process is repeated until all the remaining 9-mers are exhausted.

> *python U5.1_RemainingMinSet.py*

> *python U5.2_MinSet.py*

8. Calculate the percentage of non-redundant reduction (NRR) for the generated minimal set.

NRR is defined as the percentage of non-redundant sequences removed from a deduplicated dataset and is calculated by the use of Eq. 2:

$$Non-redundant\ reduction\ (NRR) = \frac{number\ of\ protein\ sequences\ in\ the\ deduplicated\ dataset\ (non\text{-}redundant, nr\ dataset) - number\ of\ protein\ sequences\ in\ the\ minimal\ set}{number\ of\ protein\ sequences\ in\ the\ retrieved\ dataset\ (redundant, r\ dataset)}\ X\ 100$$

For example, a deduplicated dataset of MPX with 1,245 nr sequences was used as an input for UNIQmin to generate a minimal set, which comprised of 866 sequences, resulting in a NRR (1245-866 = 379) of ~2.0% relative to the 19,423 retrieved (redundant) dataset. Thus, only less than 5% (866) of the reported monkeypox protein sequences (19,423) are required to represent the inherent viral peptidome diversity at the species rank.

Instead of running each of the individual UNIQmin scripts step-by-step as above, the user can, alternatively, execute them as a single shell script pipeline (UNIQmin.sh; available via the GitHub page) or as a Python package (available via both PyPI and the GitHub page).

Basic Protocol 3: Comparative minimal set analysis across taxonomic lineage ranks

Viruses are classified based on similarity into groups, termed as taxa. Such classification enables a survey on the extent of virus genomic diversity, possibly leading to novel insights and future research on viral origin and evolutionary relationships [30]. Starting from 2017, the hierarchy of virus taxonomy has been changed from a five-rank to 15-rank structure, including eight primary and seven derivative ranks. As of April 2022, the **I**nternational **C**ommittee on **T**axonomy of **V**iruses (ICTV) classified all known viruses into 10,434 species, 2,606 genera, 233 families, and 65 orders (https://talk.ictvonline.org/files/master-species-lists/).

In the earlier Basic Protocol 2, we described the steps to generate a minimal set of sequences and demonstrated it at the species rank for the monkeypox virus. In this Basic Protocol 3, we showcase the identification of minimal sets across higher taxonomic lineage ranks, by demonstrating for the genus and family levels. Such a comparative analysis provides for a broader understanding of minimal sets across the viral taxonomic lineage ranks, and in turn can offer a holistic understanding of sequence diversity across viral relations (from closer to distant).

1. Retrieve protein sequences of a selected virus from the publicly available database(s) of choice, across taxonomic lineage ranks (genus, and family ranks will be used as an example herein).

To demonstrate this step, MPX species will be used as the exemplar case. The Step 1 of this protocol is the same as Step 1 of the Basic Protocol 1. All reported protein sequences of MPX across the higher lineage ranks (genus and family) were collected from the NCBI Entrez Protein (NR) Database, using the respective IDs (txid: 10242 for *Orthopoxvirus* (genus rank); and 10240 for *Poxviridae* (family rank)) via the NCBI Taxonomy Browser. The protein sequences in FASTA format, with the metadata in GenPept (full) format are downloaded for each rank. As of July 2022, a total 83,088, and 163,793 viral protein sequences of genus *Orthopoxvirus* and family *Poxviridae* were extracted, respectively.

2. Deduplicate the retrieved datasets.

   See Basic Protocol 1 Step 3.

3. Implement UNIQmin to generate a minimal set of sequences for each dataset.

   See entire Basic Protocol 2.

4. Calculate the percentage of RR, NRR and total reduction (TR) for each of the minimal sets.

   Percentage of RR and NRR is to be calculated using Eq. 1 and 2, respectively (see Step 4 of Protocol 1 for RR calculation and Step 8 of Protocol 2 for NRR). The percentage of TR is to be calculated by use of Eq. 3:

$$Total\ reduction\ (TR) = \frac{number\ of\ protein\ sequences\ in\ the\ retrieved\ dataset\ (redundant,\ r\ dataset) - number\ of\ protein\ sequences\ in\ the\ minimal\ set}{number\ of\ protein\ sequences\ in\ the\ retrieved\ dataset\ (redundant,\ r\ dataset)}\ X\ 100$$

where r dataset in Eq. 3 is referred to as the retrieved dataset containing redundant sequences.

5. Compare and contrast the percentages of RR, NRR and TR for the selected virus across the taxonomic lineage ranks and analyse the relationship of all the reductions.

   As described in Basic Protocol 2, the RR for MPX (species) was copious and is expected to vary from virus to virus. The NRR (~2.0%) was relatively small and is expected to increase, transitioning from the species rank (the smallest dataset) to the family rank (the largest dataset) as the nr dataset increases to comprise more species from the different genera. Naturally, the TR is largely influenced by the RR. A summary of all the three reductions for MPX at the species, genus and family ranks is shown in Table 1 and Figure 3. The RR at the genus rank was ~81.3%, whereas the NRR was ~7.7%, with ~10.9% (9,146) of the reported genus sequences (83,088) required to represent the inherent viral peptidome diversity at the genus rank. The diversity was higher at the family rank: RR of ~78.8% and NRR of ~5.6%, with ~15.6% (25,618) of the reported family sequences (163,793) required to represent the peptidome diversity at the rank. The peptidome diversity is limited at the species rank, which is favourable for the development of intervention strategies. A comparative analysis across different viruses at multiple lineage ranks can provide valuable insights in terms of trends of how the sequence increase at the redundant and non-redundant levels contribute to the effective diversity in the minimal set.

Basic Protocol 4: Analysis of factor affecting the minimal set

A minimal set of sequences that represents the entire diversity of a given dataset can be affected by several factors, such as the i) size of the *k*-mer, ii) sequence number, iii) sequence length, and iv) sequence region (conservation level). The effect of these factors can be observed at both the reduction levels, RR and NRR. Khan (2005) had investigated the latter three factors for NRR, however, such analysis has not been done at the RR level. According to Khan (2005), increase in sequence number and length showed a general trend of requiring a larger number of sequences to represent the peptidome diversity in the minimal set, and thus, inversely proportional to the compression of the minimal set (i.e. smaller NRR with increasing sequence number and length). The selection of sequence region (conservation level) affects the minimal sequence set, where a highly diverse region, generally, results in a smaller proportion of compression because a larger number of sequences would be required to capture the entire peptidome diversity (i.e. smaller NRR with increasing sequence diversity). Additionally, it is critical to evaluate the compression of biological sequences against random sequences, reflective of background noise.

It is imperative to control for any confounders when analysing the effect of a factor to the minimal set. For example, to study the effect of *k*-mer lengths, a possible confounder would be the inherent diversity of the sequences used for the analysis, and hence, for a fair, comparable evaluation, the other factors, such as number, length, and region of the sequences would need to be controlled. Additionally, the sequences used for the comparison must be selected randomly. The confounder, region of the sequences used can be dealt by studying sequences of homogenous diversity level, which can be quantified by use of Shannon's entropy, *H*, a measure of protein sequence diversity [31]. The minimum and maximum possible entropy values for *k*-mers of nine (for immunological applications) are zero and 39, respectively [32]. However, this maximum possible entropy value is theoretical given the need to maintain conservation among related sequences for viability. According to Hu *et al.* (2013), the maximum entropy value of 9.2 was observed for envelope protein of human immunodeficiency virus type 1, clade B subtype (HIV-1 clade B), which is significantly higher than that of avian influenza [10,33,34], dengue [35] and West Nile [36] viruses. Similarly, HIV-1 clade B proteome showed the highest mean nonamer entropy value (1.9 – 4.2) compared to other viruses. As such, the sequence region diversity can be generally categorized into four groups for a factor analysis, namely highly conserved (average $H < 1$), semi-conserved ($1 \leq H < 2$), diverse ($2 \leq H < 3$) and extremely diverse ($H \geq 3$). Notably, the specific H value proposed for the grouping is arbitrary (though is reflective of the conservation levels indicated), and thus, one may consider different thresholds.

1. Select protein datasets for the different diversity level groups, defined based on Shannon's entropy value, H.

   In this protocol, for demonstration purposes, published diversity studies of three viruses, influenza A virus subtype H5N1 (H5N1) [33], dengue virus (DENV) [35] and HIV-1 clade B [11], were surveyed for proteins that met the entropy value criterion of each diversity group level. Additional criteria to manage confounders for an equitable comparison were considered for the protein selection: such as, sequence number (availability of at least 1,000 sequences) and sequence length (at least 100 amino acids long). If multiple protein datasets satisfied the above criteria for a diversity group, the dataset with average entropy closer to the mid-point of the recommended range was selected. Herein, the datasets avian H5N1 PA protein, DENV NS3 and NS2a, and HIV-1 clade B Nef were selected as test case for the four diversity categories, respectively: highly conserved, semi-conserved, diverse and extremely diverse.

2. Prepare the dataset for the selected viral proteins of interest.

   See Steps 1 and 2 of Protocol 1.

3. [Factor analysis for RR] Subsample the prepared datasets. Based on the suggested selected dataset in Step 1, a total number of four datasets should be prepared to represent the diversity spectrum.

This step may vary for each factor analysis.

(i) Sequence length

In order to study the effect of sequence length against RR, the dataset is randomly subsampled to a fixed number, 1,000 sequences and *k*-mer of nine (9) for evaluation of different sequence lengths, from 20 to 180 amino acids (aa), with a 20aa interval between the lengths, to the eventual respective protein full-length (Table 2). The sequence region for every sequence length is randomly determined. Below is an example of four datasets for the factor analysis of sequence length. Larger sequence length can be explored if the restrictions imposed by the various selection criteria to control the confounders permit.

(ii) Sequence number

Similar to the analysis of the factor, sequence length, a dataset for the four diversity level groups is randomly subsampled to a fixed, 100aa length and *k*-mer of nine (9) for evaluation of different sequence numbers, from 100 to 1,000 (Table 3). Larger number of sequences can be explored if the restrictions imposed by the various selection criteria to control the confounders permit.

4. [Factor analysis for RR] Remove duplicate sequences from all the datasets using CD-HIT.

   See Step 3 of Protocol 1.

5. [Factor analysis for RR] Calculate the percentage of RR for each dataset and compare the RR trend.

   See Step 4 of Protocol 1 for the equation. A line plot is recommended to study the RR trend.

   (i) Sequence length

A general trend of decrease in redundant reduction (RR) was observed with increase in sequence length for each of the diversity group datasets (Figure 4). The reduction was at the lowest level for the extremely diverse Nef protein of HIV-1 clade B, and the gap with the other diversity groups was closed with the full-length (FL) dataset, in particular for avian H5N1 PA (given that it had the longest full-length of 704 aa, in contrast to 202 aa for Nef), but the difference still remained significant.

Highest RR was expected for the highly conserved protein dataset, however, this was not the case for avian H5N1 PA; instead DENV NS3 behaved more like a highly conserved protein dataset, which is in agreement with the literature [35,37]. This highlights that using the average entropy to classify the proteins into the four diversity level groups may not be reflective of the actual diversity across the length of a given protein. An alternative would be to calculate the entropy for each given subsampled dataset. It should also be noted that redundancy in terms of submission of duplicates or highly similar/related sequences into database varies from virus to virus, and may not be uniform across the proteins of a virus, and thus, can impact the factor analysis evaluating RR.

(ii) Sequence number

A general trend of increase in redundant reduction (RR) was observed with an increase in sequence number (Figure 5). Such an increase in RR was at the lowest, near-uniform level for the extremely diverse Nef protein of HIV-1 clade B; however, a significant increase (~9.8%; from ~8.5% to ~18.3%) was observed for the protein when considering the larger number of sequences analysed in the full-dataset (FD) (4,350 sequences). A similar spike in RR (the highest, ~10.1%; from ~79.4% to ~89.5%) was also observed for FD of DENV NS2a, with the largest number of sequences (4,725). The DENV NS3, which had similar number of sequences in the FD (4,706) only showed a small spike (an increase of ~3.1%) as the RR was already at a peak of ~93.3% at 1000 sequences. Given the jump for the extremely diverse protein (Nef), it is recommended that such a study be carried out with even larger number of sequences. Separately, similar to the effect of length, the highly conserved protein dataset, avian H5N1 PA, was expected to have the highest RR; however, for effect of number too, it was DENV NS3 that met the expectations of a highly conserved protein, for reasons explained earlier.

6. [Factor analysis for NRR] Similar to the analysis for RR, confounders also need to be controlled for NRR.

   (i) Overlapping *k*-mer size

   The effect of *k*-mer was not explored for RR because the generation of *k*-mers is only involved in the identification of a minimal set and computation of NRR. In the case of RR, it is full-length duplicates, rather than *k*-mer duplicates that are removed for the compression. To study the effect of *k*-mer size, an nr dataset randomly subsampled to a fixed, 1,000 sequences of length 100aa would be appropriate. However, upon deduplication of the redundant, subsampled datasets (from Steps 3 of Basic Protocol 4), those that met the length criteria of 100aa contained less than 100 nr sequences. Thus, the nr dataset of the full-length was used as a starting point for this analysis (Table 4), where 300 sequences were randomly subsampled for each dataset to be used for analysis of effect of *k*-mer lengths. Window sizes from 3- to 23-mers were explored for the analysis of this factor, with the length of three representing the starting point at which significant matches become possible (basic local alignment search tool, BLAST [26,38] uses this as the default window size), up to the length of 23aa for immunological purpose. Larger *k*-mer length maybe explored if desired, with the largest recommended *k*-mer size being the length of the shortest (either full-length or partial) sized sequence in the dataset.

(ii) Sequence number

The approach is similar to the analysis of this factor for RR (four diversity level groups, length of 100aa, and *k*-mer of 9). However, similar to the analysis of *k*-mer for NRR, upon removal of duplicates, only the full-length dataset was of sufficient number of sequences for further analysis (Table 5). A length of 100aa was randomly determined for each of the nr datasets and was randomly subsampled for sequence number, from 30 to 150.

7. [Factor analysis on NRR] Calculate the percentage of NRR for each dataset and compare the AR trend.

See Step 8 of Protocol 2 for the equation. A line plot is recommended to study the AR trend.

(i) Overlapping *k*-mer size

There was a general trend of decrease in non-redundant reduction (NRR) with increase in the overlapping *k*-mer size (Figure 6). Notably, the reductions were at the lowest level for the extremely diverse dataset, Nef protein of HIV-1 clade B, with a significant gap from the other diversity group datasets. A near-saturation in the reduction, at their respective level, was observed for each of the diversity group datasets, inching closer to the largest *k*-mer analysed (23 aa).

(ii) Sequence number

A general trend of increase in NRR was observed for each of the diversity group datasets with increase in sequence number (Figure 7). The reduction levels were similar for all the diversity group datasets, except for H5N1 PA. However, it should be noted the difference was only slightly more than 3% at the highest number of sequences (150) analysed. Thus, the number of sequences analysed herein is small and the difference between the diversity group datasets is limited. This suggests that a larger number of sequences need to be analysed for a meaningful difference in trend between the diversity datasets, which was not possible with the test case proteins used herein. Nonetheless, based on the effect of sequence number with RR, the general trend for NRR is expected to be similar and maintained with increasing number of sequences, potentially expected to showcase an early large reduction (NRR) for highly conserved, semi-conserved, and diverse datasets, while a gradual one for the extremely diverse dataset at a much larger number of sequences.

It should be noted that the observation herein is in disagreement with the observation made by Khan (2005). This could be because the earlier study did not make the observation across the four conservation levels, as done herein, though the increase in the number of sequences is similar. Additionally, the increase in NRR with increase in number of unique (nr) sequences in a dataset, observed herein, can be attributed to the fact that every additional unique sequence will only contribute one or few unique *k*-mers, but many more redundant *k*-mers. This is particularly so if the sequences are of the same species.

8. [RR and NRR analysis using random sequence dataset] Generate random sequence datasets that contain different length and number of sequences using the "v_randomizer" script, publicly available at https://github.com/ChongLC/MinimalSetofViralPeptidome-UNIQmin/tree/master/randomizer. The amino acid composition of the random dataset is based on the retrieved dataset of all reported viral sequences in the NCBI Protein database (as of May 2021).

   Figure 8 clearly depicts that when compared to non-random biological sequences, no reduction was observed for random datasets, even when tested with large datasets of 100,000 sequences and of long length (3,000 aa) (Table 1 and Figure 3).

Discussion

Viral diversity, an outcome of various evolutionary forces such as mutation, recombination, and re-assortment, plays a prominent role in disease emergence and control. UNIQmin enables study of viral diversity at the protein level in the context of peptidome degeneracies. The rapid surge in sequence data has necessitated alignment-free approaches to study diversity, in particular at higher taxonomic lineages. The concept of minimal set enables the study of sequence diversity without alignment, across taxonomic lineages. This is facilitated by the tool UNIQmin, which simply takes sequence data in FASTA format and outputs a compressed minimal set. Additionally, the compression offers a much-reduced data for downstream analyses without compromising on the peptidome diversity and is particularly beneficial in the case of big-data. Herein, we provide a detailed, systematic, step-by-step protocol on data retrieval from public databases and preparation (BP1), generation of a minimal set (BP2), its utility across taxonomic lineages (BP3), and investigating the effects of various factors affecting the minimal set (BP4). UNIQmin is applicable to viral and possibly other pathogenic microorganisms, with the possibility of dissecting the diversity spatio-temporally to allow for comparative analyses.

Herein, we also performed alignment-free study of monkeypox viral diversity (species level) and its higher taxonomic lineage ranks (genus and family). Only less than 5% of the reported monkeypox protein sequences are required to represent the inherent viral peptidome diversity at the species rank, which increased to less than 16% at the family rank. The findings have important implications in the design of vaccines, drugs, and diagnostics, as only a small number of sequences are required for coverage of the viral diversity.

Data Availability

The tool and the data underlying this article are available in GitHub at https://github.com/ChongLC/MinimalSetofViralPeptidome-UNIQmin and https://github.com/ChongLC/UNIQmin_PublicationData, respectively. The datasets were derived from sources in the public domain, namely NCBI Protein database (https://www.ncbi.nlm.nih.gov/protein). The tool is also available at PyPI (https://pypi.org/project/uniqmin/).

Key points

- UNIQmin offers an alignment-free approach for the study of viral sequence diversity at any given rank of taxonomy lineage and is big data ready
- The tool performs an exhaustive search to determine the minimal set of sequences required to capture the peptidome diversity within a given dataset.
- As an alignment-free approach removes various restrictions, otherwise posed by an alignment-dependent approach for the study of sequence diversity.
- The problem-solving protocol and case study described herein facilitate systematic diversity studies across viral taxonomic lineages.
- Alignment-free sequence diversity study of monkeypox virus across its taxonomy lineage (species, genus and family) revealed that only less than 5% of the reported monkeypox protein sequences are required to represent the inherent viral peptidome diversity at the species rank, which increased to ~16% at the family rank. The low peptidome diversity, in particular at the species rank, is favourable for the development of intervention strategies.

Funding

AMK was supported by Perdana University, Malaysia, Bezmialem Vakif University, Turkey, and The Scientific and Technological Research Council of Turkey (TÜBİTAK). This publication/paper has been produced benefiting from the 2232 International Fellowship for Outstanding Researchers Program of TÜBİTAK (Project No: 118C314). However, the entire responsibility of the publication/paper belongs to the owner of the publication/paper. The financial support received from TÜBİTAK does not mean that the content of the publication is approved in a scientific sense by TÜBİTAK.

Acknowledgements

We gratefully acknowledge the support of Mr. Lim Wei Lun in maintaining UNIQmin.

Competing interests

LCC and AMK declare no conflicts of interest.

References

1. Domingo E, Holland JJ. RNA Virus mutations and fitness for survival. Annu. Rev. Microbiol. 1997; 51:151–178

2. Peck KM, Lauring AS. Complexities of Viral Mutation Rates. J. Virol. 2018; 92:

3. Domingo E, Perales C. Viral quasispecies. PLOS Genet. 2019; 15:e1008271

4. Volkov I, Pepin KM, Lloyd-Smith JO, et al. Synthesizing within-host and population-level selective pressures on viral populations: the impact of adaptive immunity on viral immune escape. J. R. Soc. Interface 2010; 7:1311–1318

5. Thompson JD, Linard B, Lecompte O, et al. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS One 2011; 6:e18093

6. Batzloff M, Yan H, Davies M, et al. Preclinical evaluation of a vaccine based on conserved region of M protein that prevents group A streptococcal infection. Indian J. Med. Res. 2004; 119 Suppl:104–7

7. Haubold B, Reed FA, Pfaffelhuber P. Alignment-free estimation of nucleotide diversity. Bioinformatics 2011; 27:449–455

8. Sitbon E, Pietrokovski S. Occurrence of protein structure elements in conserved sequence regions. BMC Struct. Biol. 2007; 7:3

9. Yang ZF, Mok CKP, Liu XQ, et al. Clinical, Virological and Immunological Features from Patients Infected with Re-Emergent Avian-Origin Human H7N9 Influenza Disease of Varying Severity in Guangdong Province. PLoS One 2015; 10:e0117846

10. Heiny AT, Miotto O, Srinivasan KN, et al. Evolutionarily Conserved Protein Sequences of Influenza A Viruses, Avian and Human, as Vaccine Targets. PLoS One 2007; 2:e1190

11. Hu Y, Tan PT, Tan TW, et al. Dissecting the Dynamics of HIV-1 Protein Sequence Diversity. PLoS One 2013; 8:e59994

12. Chong LC, Lim WL, Ban KHK, et al. An Alignment-Independent Approach for the Study of Viral Sequence Diversity at Any Given Rank of Taxonomy Lineage. Biology. 2021; 10:853

13. Zielezinski A, Vinga S, Almeida J, et al. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017; 18:186

14. Ren J, Bai X, Lu YY, et al. Alignment-Free Sequence Analysis and Applications. Annu. Rev. Biomed. Data Sci. 2018; 1:93–114

15. Khan AM, Heiny A, Lee KX, et al. Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus. BMC Bioinformatics 2006; 7:S4

16. Khan AM. Mapping targets of immune responses in complete Dengue viral genomes. Sch. Repos. 2005

17. Heiny AT. Characterizing evolutionarily conserved influenza A virus sequences as vaccine targets. 2009

18. Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021; 49:D10–D17

19. Hatcher EL, Zhdanov SA, Bao Y, et al. Virus Variation Resource – improved response to emergent viral outbreaks. Nucleic Acids Res. 2017; 45:D482–D490

20. Pickett BE, Sadat EL, Zhang Y, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 2012; 40:D593–D598

21. Khare S, Gurry C, Freitas L, et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021; 3:1049–1051

22. Stoesser G, Griffith M, Griffith OL. HIV Sequence Database. Dict. Bioinforma. Comput. Biol. 2004

23. Neogi SG, Vasilis P, Krestyaninova M, et al. MoDa-A Data Warehouse for Multi-"Omics" Data. J. Data Mining Genomics Proteomics 2013; 04:145

24. Schonbach C, Kowalski-Saunders P, Brusic V. Data warehousing in molecular biology. Brief. Bioinform. 2000; 1:190–198

25. Koh JLY, Lee ML, Khan AM, et al. Duplicate Detection in Biological Data using Association Rule Mining. Proc. Second Eur. Work. Data Min. Text Min. Bioinforma. 2003; 35–41

26. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J. Mol. Biol. 1990;

27. Bateman A, Martin M-J, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49:D480–D489

28. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; 44:D733–D745

29. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22:1658–1659

30. Simmonds P, Adams MJ, Benkő M, et al. Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 2017; 15:161–168

31. C. E. Shannon. A Mathematical Theory of Communication. Bell Syst. Tech. J. 1948

32. Khan AM, Hu Y, Miotto O, et al. Analysis of viral diversity for vaccine target discovery. BMC Med. Genomics 2017; 10:78

33. Abd Raman HS, Tan S, August JT, et al. Dynamics of Influenza A (H5N1) virus protein sequence diversity. PeerJ 2020; 7:e7954

34. Tan S, Sjaugi MF, Fong SC, et al. Avian Influenza H7N9 Virus Adaptation to Human Hosts. Viruses 2021; 13:871

35. Chong LC, Khan AM. Identification of highly conserved, serotype-specific dengue virus sequences: implications for vaccine design. BMC Genomics 2019; 20:921

36. Koo QY, Khan AM, Jung K-OO, et al. Conservation and variability of West Nile virus proteins. PLoS One 2009; 4:e5352

37. Khan AM, Miotto O, Nascimento EJM, et al. Conservation and variability of dengue virus proteins: Implications for vaccine design. PLoS Negl Trop Dis 2008; 2:e272

38. Mahram A, Herbordt MC. Fast and accurate NCBI BLASTP: Acceleration with multiphase FPGA-based prefiltering. Proc. 24th ACM Int. Conf. Supercomput. - ICS '10 2010; 73

Table and Figure Legends

Table 1. A summary of all the three reductions (redundant (RR), non-redundant (NRR), and total (TR) reductions) across the taxonomic lineages ranks of monkeypox virus, namely species, genus, and family.

Taxonomic lineage rank	**Number of protein sequences** (r dataset | nr dataset | minimal dataset)	**Percentage of reduction** (%; 1 d.p.)		
		RR	**NRR**	**TR**#
Species –*Monkeypox virus* *	19,423 | 1,245 | 866	93.6	2.0	95.5
Genus – *Orthopoxvirus*	83,088 | 15,523 | 9,146	81.3	7.7	89.1
Family – *Poxviridae*	163,793 | 34,782 | 25,618	78.8	5.6	84.4

Abbreviation: * MPX; # The addition of RR and NRR may not equal to TR due to rounding issue; r, redundant; nr, non-redundant; d.p., decimal place

Table 2. Datasets for analysis of effect of sequence length, evaluating RR compression.

Diversity spectrum | Virus and selected protein (full-length in aa)	**Sequence length (aa)**									
	20	**40**	**60**	**80**	**100**	**120**	**140**	**160**	**180**	**FL**
	Starting amino acid position in the protein (randomly determined, where possible) for each subsampled dataset of a fixed number, 1,000 sequences, and *k*-mer of 9, but of varying length (starting from 20aa to FL)									
Highly conserved ($H < 1$) | Avian H5N1 PA (704 aa)	552	22	78	124	226	436	123	485	168	1
Semi-conserved ($1 \leq H < 2$) | DENV NS3 (618 aa)	467	13	106	288	429	221	200	198	301	1
Diverse ($2 \leq H < 3$) | DENV NS2a (218 aa)	67	147	11	39	30	45	78	47	3	1
Extremely diverse ($H \geq 3$) | HIV-1 clade B Nef (202 aa)	162	33	57	89	5	4	17	18	9	1

Abbreviation: aa, amino acid; FL, full length; H5N1, influenza A virus subtype H5N1; DENV, dengue virus (1-4 serotypes); HIV-1 clade B, human immunodeficiency virus type 1 clade B subtype

Table 3. Datasets for analysis of effect of sequence number, evaluating RR compression.

Diversity spectrum | Virus and selected protein | Total number of sequences in the redundant dataset	**Starting amino acid position in the protein (randomly determined) for each subsampled dataset of a fixed length, 100aa, and *k*-mer of 9, but of varying number (starting from 100 to 1,000 sequences)**
Highly conserved (H < 1) | Avian H5N1 PA | 1,504	78
Semi-conserved (1 ≤ H < 2) | DENV NS3 | 4,706	480
Diverse (2 ≤ H < 3) | DENV NS2a | 4,725	8
Extremely diverse (H ≥ 3) | HIV-1 clade B Nef | 4,350	45

Abbreviation: H5N1, influenza A virus subtype H5N1; DENV, dengue virus (1-4 serotypes); HIV-1 clade B, human immunodeficiency virus type 1 clade B subtype

Table 4. Datasets for analysis of the effect of *k*-mer size, evaluating NRR compression.

Diversity spectrum | Virus and selected protein | Longest analysed sequence length	**Total number of non-redundant sequences, which allowed for random subsampling of a fixed, 300 sequences from a randomly determined region of length, 100aa to evaluate the effect of *k*-mers, starting from 3- to 23-mers**
Highly conserved (H < 1) | Avian H5N1 PA | 704 aa	699
Semi-conserved (1 ≤ H < 2) | DENV NS3 | 618 aa	361
Diverse (2 ≤ H < 3) | DENV NS2a | 218 aa	365
Extremely diverse (H ≥ 3) | HIV-1 clade B Nef | 202 aa	999

Abbreviation: H5N1, influenza A virus subtype H5N1; DENV, dengue virus (1-4 serotypes); HIV-1 clade B, human immunodeficiency virus type 1 clade B subtype

Table 5. Datasets for analysis of effect of sequence number, evaluating NRR compression.

Diversity spectrum | Virus and selected protein	**Total number of non-redundant sequences, which allowed for random subsampling of different sequence number (30 to 150) from a randomly determined region of a fixed length, 100aa and *k*-mer of 9**
Highly conserved (H < 1) | Avian H5N1 PA	232
Semi-conserved (1 ≤ H < 2) | DENV NS3	169
Diverse (2 ≤ H < 3) | DENV NS2a	496
Extremely diverse (H ≥ 3) | HIV-1 clade B Nef	3,556

Abbreviation: H5N1, influenza A virus subtype H5N1; DENV, dengue virus (1-4 serotypes); HIV-1 clade B, human immunodeficiency virus type 1 clade B subtype

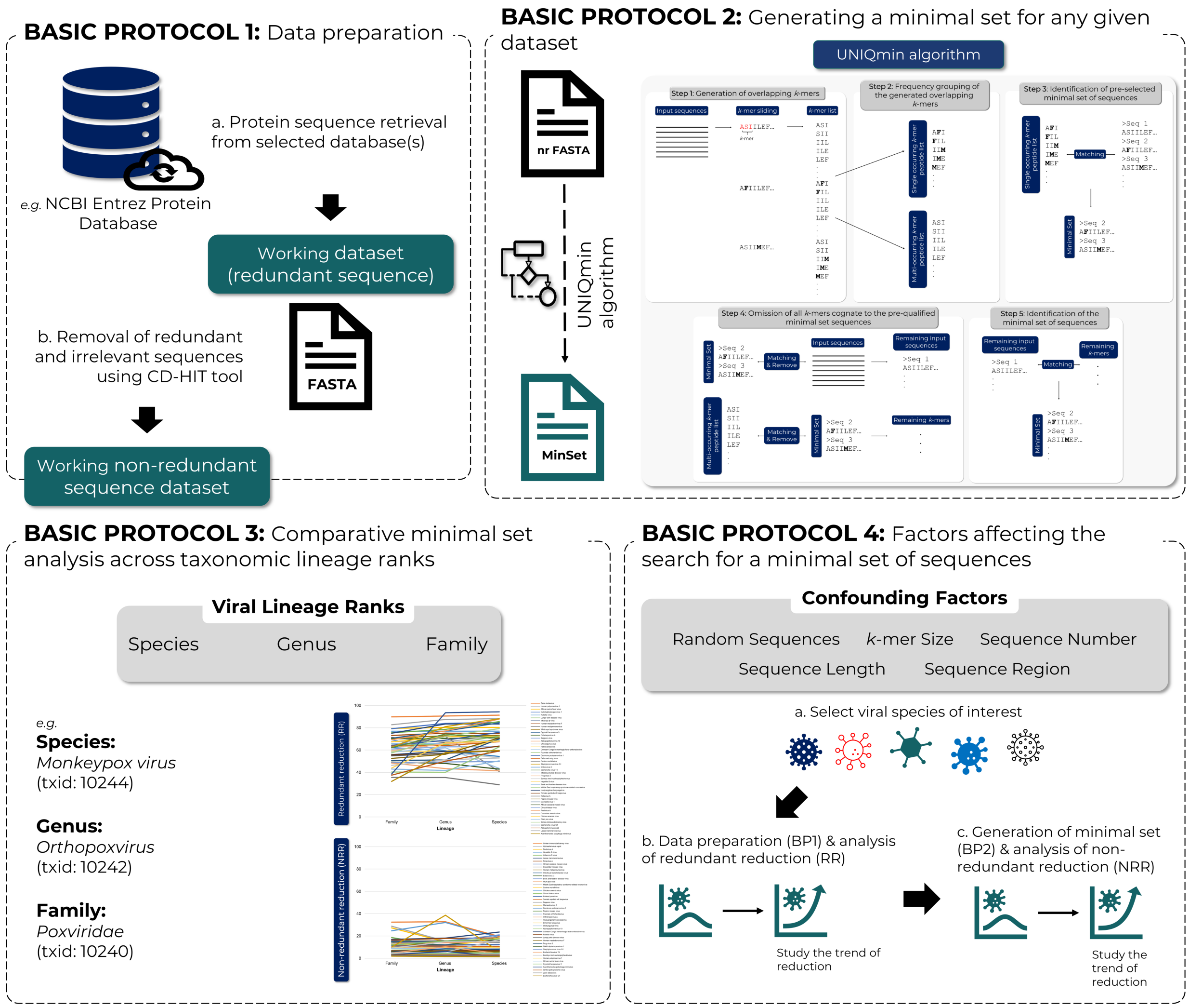

Figure 1. An overview of the basic protocols to map and study the minimal set of a viral peptidome. Protein sequences of interest are retrieved from a selected database as the working dataset. The dataset is removed of redundant and irrelevant sequences by use of CD-HIT, resulting in a non-redundant (nr) sequence dataset. UNIQmin is then applied to the nr dataset to generate a minimal set of sequences. Minimal set can be generated across taxonomic lineage ranks for comparative analysis, and the effect of factors (random sequences, size of *k*-mer, and number, length and region of sequences) affecting the minimal set can be delineated.

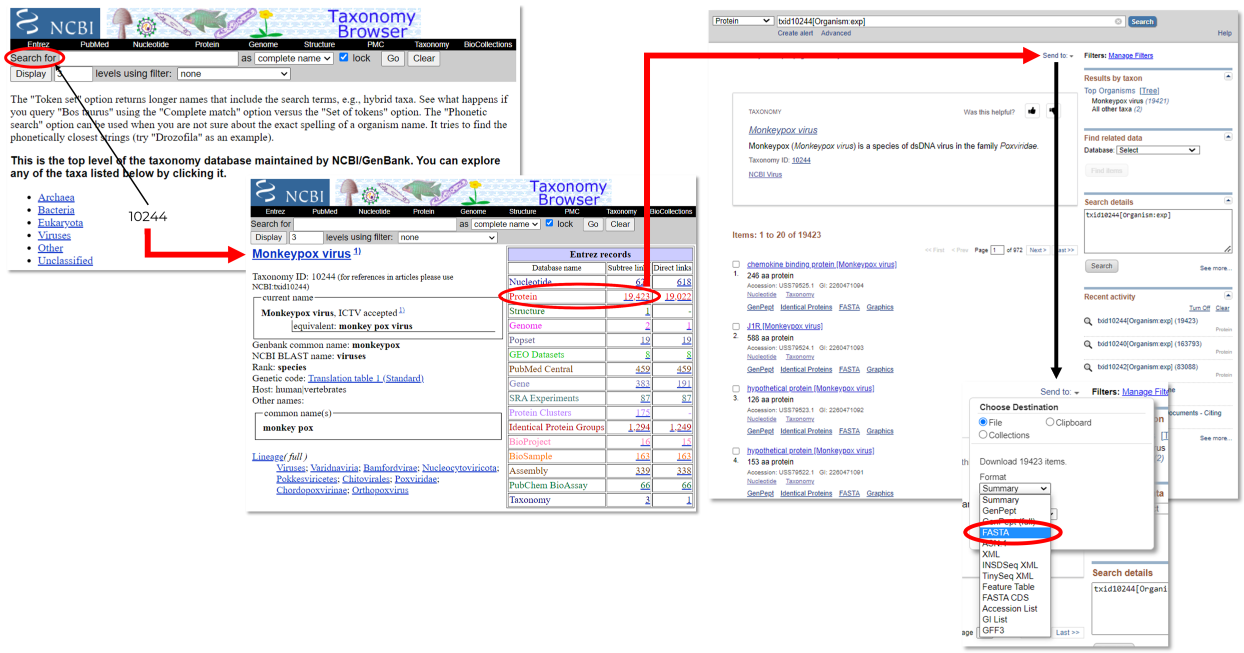

Figure 2. Data retrieval from the National Center for Biotechnology Information (NCBI) Entrez Protein (NR) Database for all reported *Monkeypox virus* protein sequences via the NCBI Taxonomy Browser, using the species taxonomy identifier (txid) "10244".

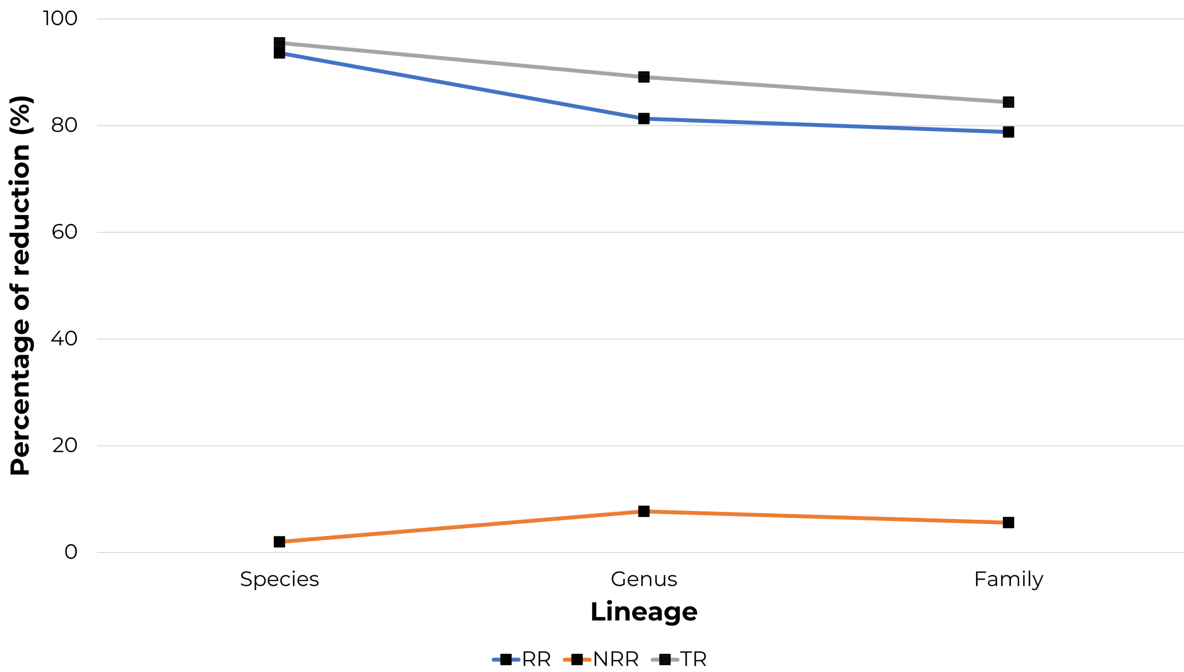

Figure 3. Percentage of reductions (%) for *Monkeypox virus* (MPX) across the viral taxonomic lineage ranks, namely species, genus, and family. Abbreviations: RR, redundant reduction, NRR, non-redundant reduction, TR, total reduction.

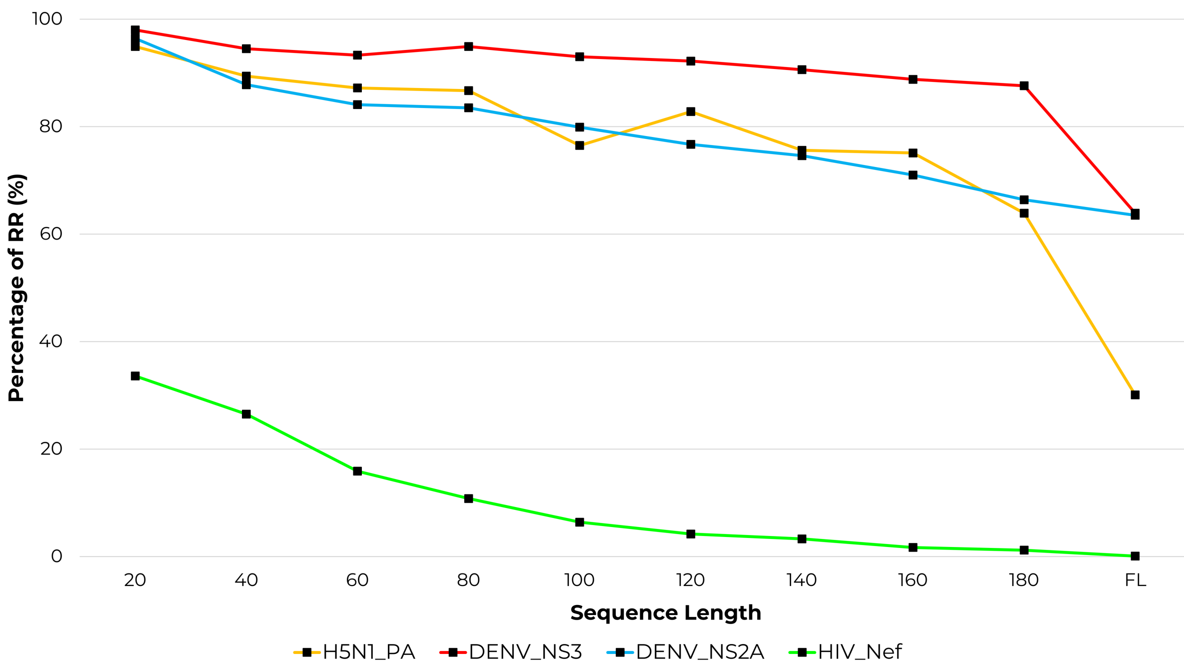

Figure 4. Effect of sequence length (aa) to redundant reduction (RR). Possible confounding factors controlled: sequence number (1,000), window size of *k*-mer (9), and diversity level groups (across diversity spectrum). Abbreviation: FL, full length; H5N1_PA, avian influenza A virus subtype H5N1 PA protein; DENV_NS3, dengue virus (1-4 serotypes) NS3 protein; DENV_NS2a, dengue virus (1-4 serotypes) NS2a protein; HIV_Nef, human immunodeficiency virus type 1 clade B subtype Nef protein.

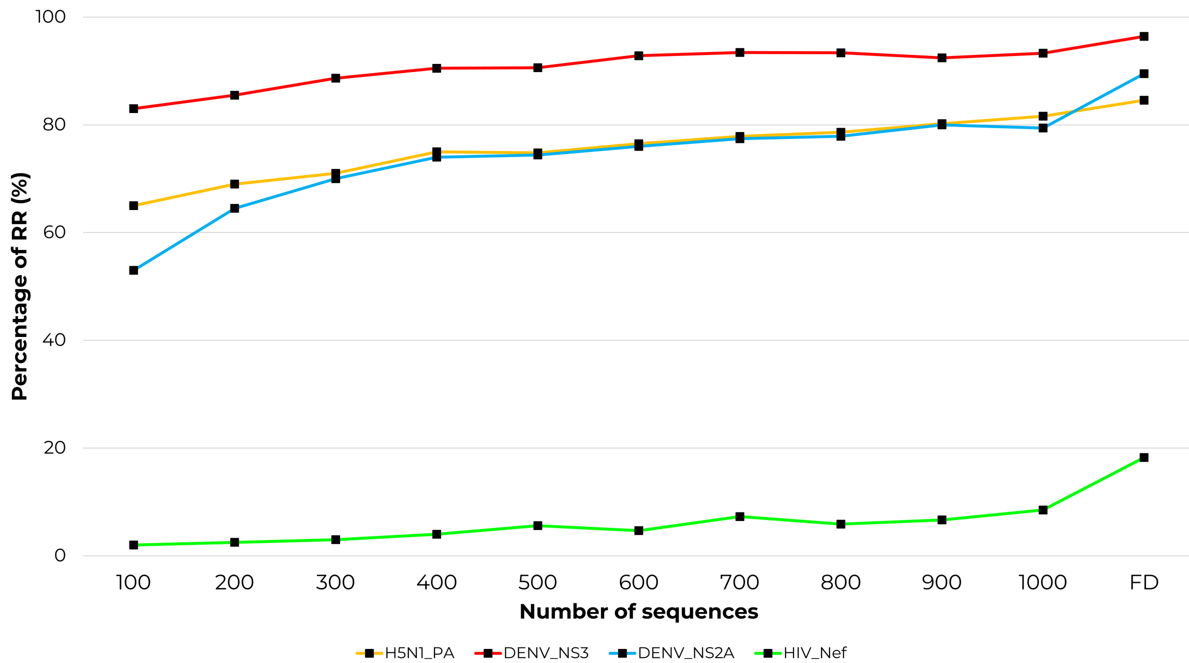

Figure 5. Effect of number of sequences to redundant reduction (RR). Possible confounding factors controlled: sequence length (100aa), window size of *k*-mer (9), and diversity level groups (diversity spectrum). Abbreviation: FD, full dataset analysed; H5N1_PA, avian influenza A virus subtype H5N1 PA protein; DENV_NS3, dengue virus (1-4 serotypes) NS3 protein; DENV_NS2a, dengue virus (1-4 serotypes) NS2a protein; HIV_Nef, human immunodeficiency virus type 1 clade B subtype Nef protein.

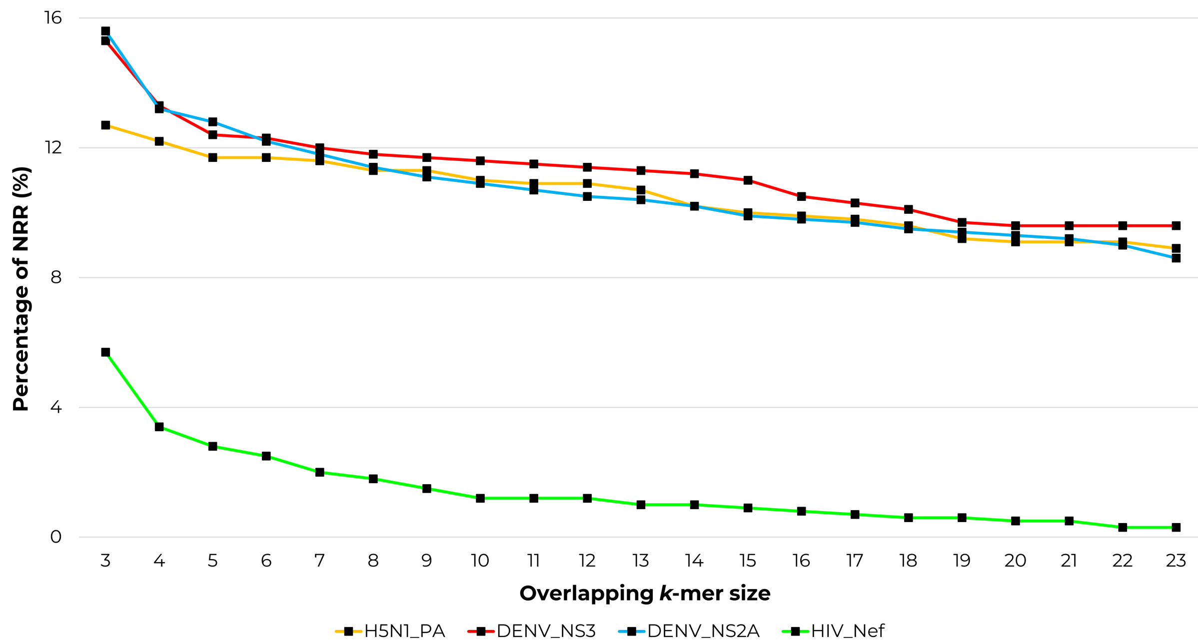

Figure 6. Effect of overlapping *k*-mer size to non-redundant reduction (NRR). Possible confounding factors controlled: sequence length (100aa), sequence number (300), diversity level groups (across diversity spectrum). Abbreviation: H5N1_PA, avian influenza A virus subtype H5N1 PA protein; DENV_NS3, dengue virus (1-4 serotypes) NS3 protein; DENV_NS2a, dengue virus (1-4 serotypes) NS2a protein; HIV_Nef, human immunodeficiency virus type 1 clade B subtype Nef protein.

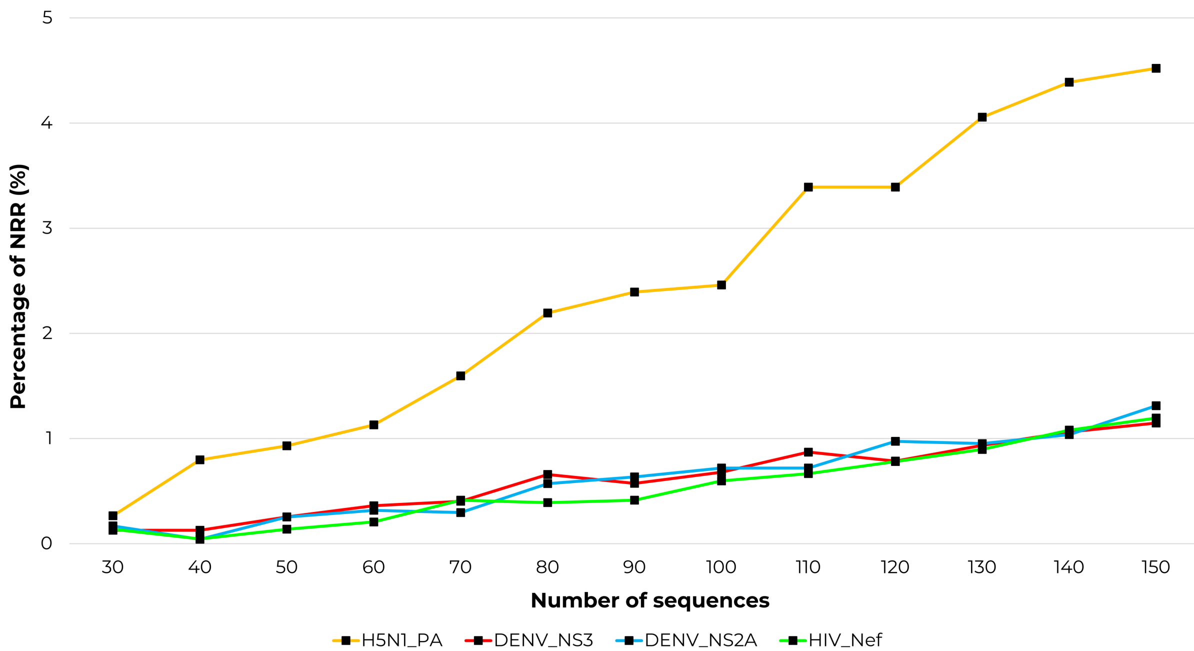

Figure 7. Effect of sequence number to non-redundant reduction (NRR). Possible confounding factors controlled: sequence length (100aa), *k*-mer (9), diversity level groups (across diversity spectrum). Abbreviation: H5N1_PA, avian influenza A virus subtype H5N1 PA protein; DENV_NS3, dengue virus (1-4 serotypes) NS3 protein; DENV_NS2a, dengue virus (1-4 serotypes) NS2a protein; HIV_Nef, human immunodeficiency virus type 1 clade B subtype Nef protein.

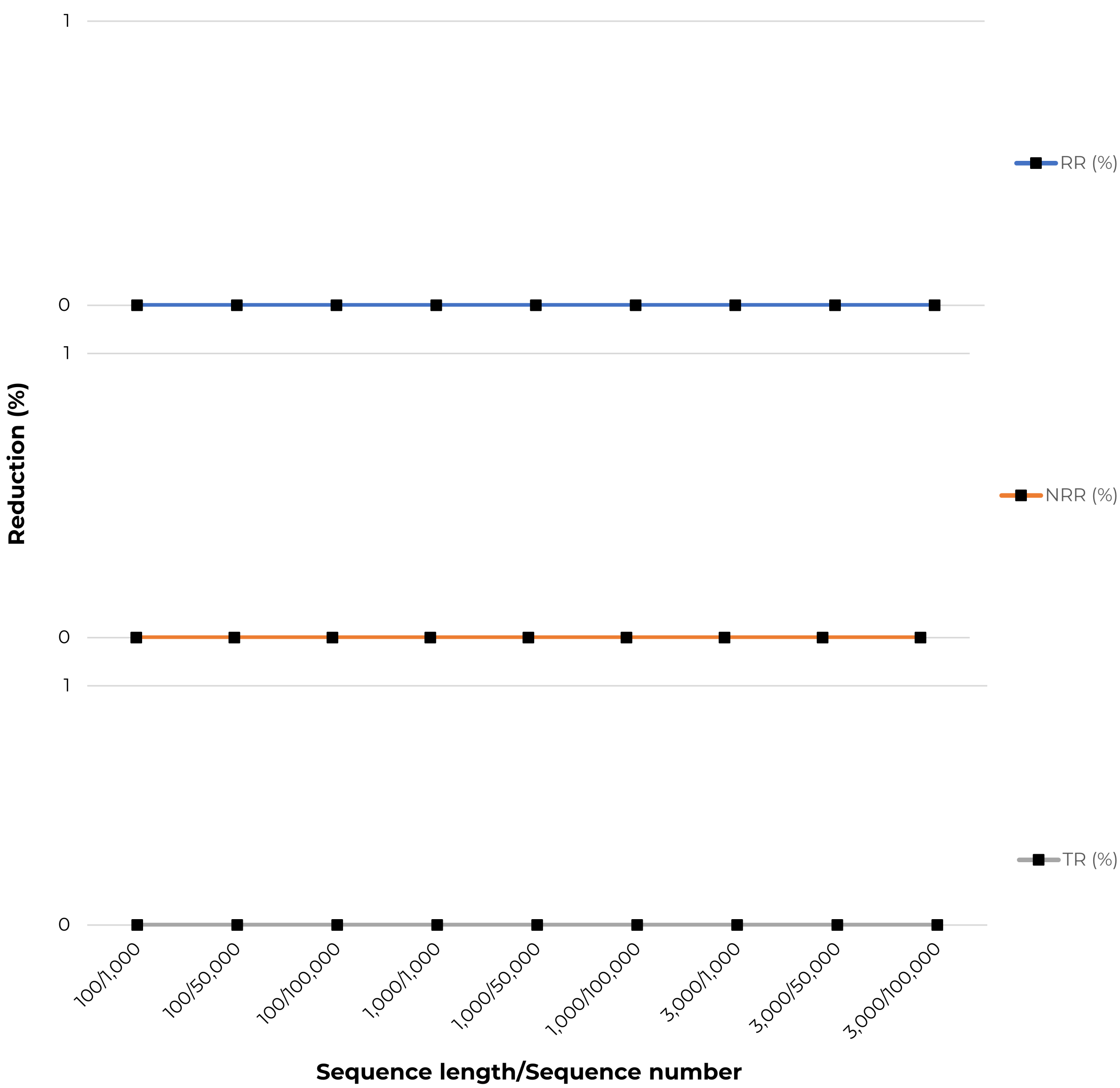

Figure 8. Minimal set for random sequences. Abbreviation: redundant reduction (RR), non-redundant reduction (NRR) and total reduction (TR).