Effective Primer Design for Genotype and Subtype Detection of Highly Divergent Viruses in Large Scale Genome Datasets

Burak Demiralay[1*] and Tolga Can[2*]

[1*]Department of Health Informatics, Informatics Institute, Middle East Technical University, Dumlupınar Bulvarı No 1, Çankaya, 06800, Ankara, Türkiye.

[2*]Department of Computer Science, Colorado School of Mines, 1501 Illionis St, Golden, 80401, CO, USA.

*Corresponding author(s). E-mail(s): burak.demiralay@metu.edu.tr; tolgacan@mines.edu;

Abstract

Identification of microorganisms in a biological sample is a crucial step in diagnostics, pathogen screening, biomedical research, evolutionary studies, agriculture, and biological threat assessment. While progress has been made in studying larger organisms, there is a need for an efficient and scalable method that can handle thousands of whole genomes for organisms with high mutation rates and genetic diversity such as single stranded viruses. In this study, we developed a novel method to identify subsequences for detection of a given species/subspecies in a (meta)genomic sample using the Polymerase Chain Reaction (PCR) method. Species detection in any analysis depends highly on the measurement method and since thermodynamic interactions are critical in PCR, thermodynamics is the main driving force in the proposed methodology. Our method is parallelized in multiple steps and involves extracting all oligonucleotides from target genomes. We then locate the target sites for each oligonucleotide using the constructed suffix array and local alignment followed by thermodynamic interaction assessment. An important requirement for subspecies identification is to avoid amplifying a non-target set of genomes and our method addresses this. We applied our method to three highly divergent viruses; 1) Hepatitis C virus (HCV),

where the subtypes differ in 31%-33% of nucleotide sites on average, 2) Human immunodeficiency virus (HIV), for which, 25-35% between-subtype and 15-20% within-subtype variation is observed, and 3) the Dengue virus, whose respective genomes (only DENV 1–4) share 60% sequence identity to each other. Using our method, we were able to select oligonucleotides that can identify *in silico* 99.9% of 1657 HCV genomes, 99.7% of 11838 HIV genomes, and 95.4% of 4016 Dengue genomes. We also show subspecies identification on genotypes 1-6 of HCV and genotypes 1-4 of the Dengue virus with more than 99.5% true positive and less than 0.05% false positive rate, on average. None of the state-of-the-art methods can produce oligonucleotides with this specificity and sensitivity on highly divergent viral genomes like the ones studied in this article.

Keywords: genome analysis, primer and probe design for PCR, viral diagnostics

1 Introduction

DNA signatures are sequences that can distinguish a group of interest from a background group of sequences. While differences in more conserved regions, such as rRNA sites, are still used for species identification, species-specific oligonucleotide strings can be found anywhere in the genome and can serve as better discriminators. Therefore, processing the entire genome is crucial to identifying species-specific oligonucleotide sequences. The ability to distinguish one organism or subtype from others has various useful applications in public health, biomedical research, agriculture, and evolutionary research, as well as in combating bioterrorism; therefore, there is a large body of research in this area. Notable studies are discussed below.

In 2001, Li and Stormo proposed a method for selecting optimal DNA oligos for gene expression arrays [1]. The proposed algorithm involved creating a suffix array of coding sequences, choosing probe candidates from every gene based on sequence features, and determining the positions of matched sequences in all genes. The free energies (dG) and the melting temperatures (Tm) of potential candidate sequences are calculated and the most discriminating probes are selected based on free energy. Same year, for the problem of PCR amplification of distantly related species, Rose

et al. developed a unique approach called CODEHOP [2]. In CODEHOP, amino acid sequences are aligned and motif identification methods are applied, taking codon usage preferences into account; thereby, encoding the conserved amino acid sequences, to amplify distantly related species. After motifs are found, amino acid blocks are turned into degenerate primers. In 2004, Gadberry *et al.* developed a tool called Primaclade for identifying conserved PCR primers across multiple species [3]. In Primaclade, a multiple sequence alignment (MSA) of target sequences is constructed and individual oligonucleotides are compared to the consensus sequence, scored in terms of degeneracy. Because MSA does not work well with high number of inputs, preliminary clustering of similar sequences is suggested. In 2006, Jabado *et al.* proposed a method for designing degenerate primers for viruses [4]. To decrease negative implications of degeneracy, they modeled the problem as a set cover problem and sought for the minimum number of primer sets. Their main algorithm is to extract small subalignments of the multiple sequence alignment to be used as primers and build a phylogenetic tree. Consensus sequences for every branch are identified and scored against all others for finding the minimum number of primers to amplify all sequences. In 2009, Duitama *et al.* developed PrimerHunter, a tool that differentiates target and non-target virus sequences [5]. They emphasize that degenerate primer approaches ignore primer specificity which prevents their use for direct viral subtyping assays. PrimerHunter exhaustively generates primers from target sequences by searching and counting user specified seed sequences. It then performs filtering by several constraints including the melting temperature (Tm). Their method also used a minimal set cover approach when a single primer set could not amplify all sequences. In 2010, Vijaya Satya *et al.* introduced the Tool for PCR Signature Identification (TOPSI), a pipeline for discovering real-time PCR signatures [6]. TOPSI uses pairwise alignments, extracts common sequences among target genomes, incorporates various constraints to generate candidate primers and probes, and uses BLAST against non-target genomes

for specificity analysis. Because the tool needs to find conserved regions to generate oligonucleotides, it does work well on bacteria but not on highly variable viruses. In 2012, Hysom *et al.* proposed a method that extracts k-length oligonucleotides from all targets and counts them [7]. Their method picks the most conserved *k-mers* and realigns them to targets while allowing mismatches. Other primer pairs for remaining targets are then found iteratively. In 2014, Lee and Sheu proposed an algorithm and employed a divide-and-conquer strategy and a parallel signature discovery approach [8]. They define a signature, a fixed length l with allowed mismatches d and this (l, d) pattern is required to occur only once. A given genome is recursively divided into pieces until full pattern search can be run directly on that piece. After the patterns are found in each piece, the pieces are merged and patterns that are found on any other piece are eliminated. In 2017, Marinier *et al.* developed Neptune for identifying differentially abundant genomic content in bacterial populations [9]. It uses fixed size *k-mer* matching with a probabilistic model to find the best cardinality of k. BLAST is used for further refinement. In 2019, Karim *et al.* developed a primer design pipeline, Uniqprimer, to distinguish target genomes from non-target genomes [10]. A single target genome is first aligned to all non-target genomes and non-aligned regions are extracted. Then, these regions are aligned to another target genome and common regions are iteratively aligned to all target genomes. Primers are designed from these conserved regions. In 2022, Metsky *et al.* developed a pipeline, where they use multiple sequence alignment and then, with neural networks, solve a complicated scoring function for activity of a probe for CRISPR-based virus identification [11].

1.1 The Rationale Behind Our Approach

We firmly believe that the hybridization efficiency of two DNA strands should be evaluated exclusively based on thermodynamic principles, as these govern the interactions of DNA sequences in laboratory settings, rather than relying on fixed sequence-based

parameters such as the number of mismatches. Thermodynamics governs interactions of sequences in a laboratory setting. Calculation of Tm of an oligonucleotide with a different non-complementary oligonucleotide is a complex procedure, and an optimization problem must be solved recursively with fractional programming where the set of all possible alignments of two sequences, and the enthalpy and entropy differences of the corresponding chemical reactions are the variables [12]. This calculation is extremely slow and is the bottleneck of any method. However, with careful planning, this bottleneck can be overcome and sequence similarity, either suffix array or k-mer based, must only be an intermediate step to reduce running time because of the complexity of thermodynamic analysis; therefore, a small number of non-stringent parameters must be used.

We emphasize that oligonucleotide design based on the number of mismatches may be very misleading. On the left pane of Figure 1, we show an example of this condition.

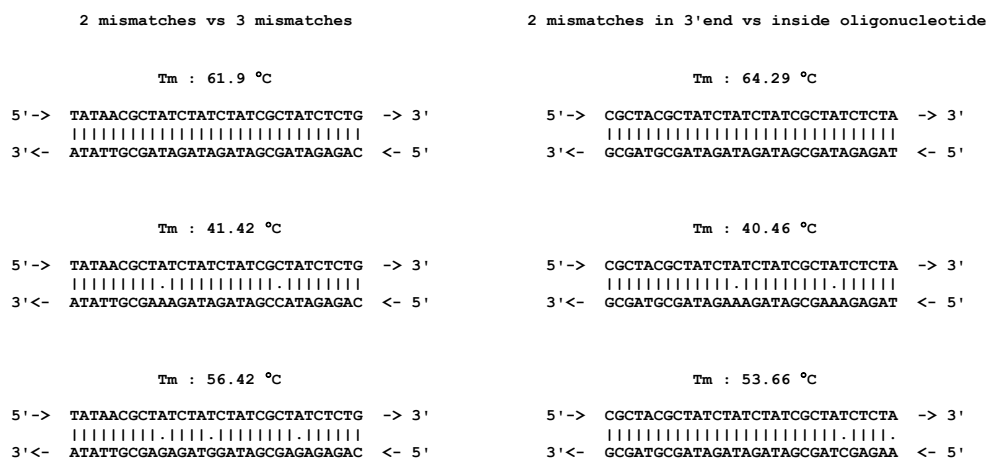

Fig. 1 Melting temperatures for different number of mismatches. Tm indicates the melting temperature.

Figure 1 shows that an oligonucleotide's interaction with its complementary sequence has a much higher binding affinity when there are two mismatches compared to three mismatches, with a 15 °C difference. A random 25bp oligonucleotide, when complemented with three mismatches, has 0.086 probability of having a higher Tm than five mismatches and this probability raises to 0.203 when the temperature difference is kept within five degrees.

The 3′ end of oligonucleotide conservation is also used by various methods to reduce the large search space. The rationale for this heuristic is that polymerase enzymes add new bases from the 3′ end of the oligonucleotide, and it is possible that stable binding of the 3′ end of the oligonucleotide may be enough for polymerases to start extension even though the rest of the oligonucleotide binds weakly [13]. Specifically, for a random 25bp oligonucleotide, the interaction with its complementary sequence, with a mutation outside the first five bases of the 3′ end, has about 0.025 probability of having a higher Tm than the interaction with its complementary sequence with a mutation in the last base of the 3′ end. However, this probability increases to 0.30 when the Tm difference is kept at 5 °C. This would lead to generating more false positives in differentiation studies and could reduce true positive rate. We argue that any similarity-based heuristics cannot capture these binding affinities. The right pane of Figure 1 shows an example where an oligonucleotide with mutations in the 3′ end is more favorable.

Our method involves thermodynamic analysis around locally aligned regions in whole genomes while parameterizing key variables, and we believe that our method of local alignments based on more lenient sequence similarity is devoid of these shortcomings, as supported by the results presented in this study. In addition, accurate multiple sequence alignment of whole genomes is computationally expensive, and common/consensus region approaches do not work on divergent sequences, as we also show in the results section. Degenerate primers work on amplifying common regions at the

cost of introducing false positives in subtype detection. We believe our method also addresses these shortcomings.

1.2 The Proposed Method

Our problem involves identifying primer-probe sets for PCR that can specifically bind and amplify each genome in a given set of target genomes, while not binding to any genomes in another set of background genomes, thereby identifying the target organisms.

Figure 2 shows the steps of our proposed solution, which are described below.

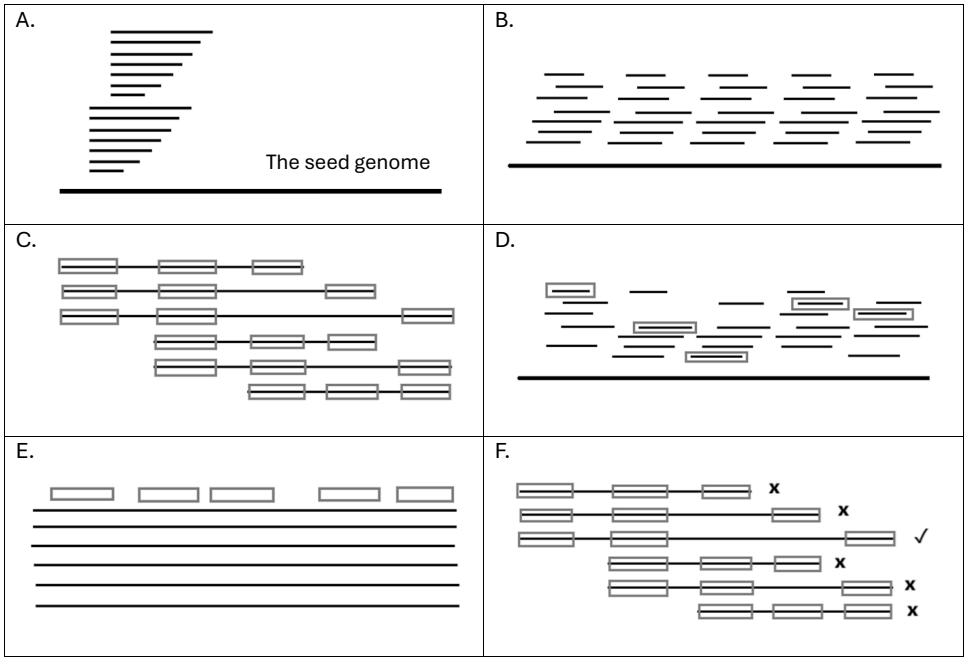

Fig. 2 **A.** Generate all possible primers and probes from the seed genome with respect to constraints defined by laboratory conditions. **B.** Group oligonucleotides according to their start and end positions. **C.** Considering all oligonucleotides, form all possible amplicons on the seed genome with respect to length, temperature difference, and interaction constraints. On the image, the flow is simplified; the primers and the probes come from different sets. **D.** Discard oligonucleotides that do not appear in an amplicon. Randomly select k representatives from every group of valid oligonucleotides. **E.** Locally align every oligonucleotide to every other genome with lenient parameters and find hit locations. **F.** Compute interaction Tms of oligonucleotides and reconstruct amplicons on every other genome. After these steps, as a final step, post filter the results with respect to target specificity and sensitivity.

The input genomes can be filtered based on the number of non-ACGT bases they contain or based on the length of the input genomes mainly to ensure data quality. After filtering, an optional starting step is to generate a consensus genome from the common regions of a given subset of the genome set. In this step, common regions are extracted and reassembled using Mummer4 [14]. This process is repeated using stepwise pairwise alignment. The stage of finding common regions among all or a subset of target genomes may be preferred to shorten the running time for organisms with larger genomes or organisms with low mutation rates. For viruses with high mutation rates, it would be more suitable to choose only a single genome for execution, so the consensus genome would be the single provided genome.

The next step is extracting oligonucleotides that satisfy the given constraints from input genomes. This stage is an oligonucleotide scan using a sliding window. We scan the genome in both strands because oligonucleotides extracted from one strand may fit given constraints while complementary oligonucleotides may not. The variables that we use in this stage are: 1) acceptable melting temperature ranges for primers and probes, 2) length ranges of oligonucleotides, and 3) laboratory variables such as concentrations of monovalent cations, divalent cations, the primer, the probe, and the DNA.

We then generate the amplicons formed by the oligonucleotides identified in the previous stage. Generally, shorter than 300 base pair amplicons in PCR are amplified more easily because common DNA Polymerases tend to hang and fall from longer stretches of DNA. So, before querying the presence of oligonucleotides on other genomes, we make sure that those oligonucleotides form valid amplicons that can be validated in laboratory conditions, and we discard oligonucleotides that are not used in any formed amplicon. Since the location and length of each oligonucleotide are known, this step is not computationally intensive. The variables that we use in this stage are the length range of amplicons, the desired minimum Tm difference between

primers and probes, and the maximum allowed Tm difference between two primers. Note that the number of sequences to be extracted from the first input genome could be extremely high. A 9000-base pair HIV virus yields approximately 400,000 short oligonucleotides that can be used as primers or probes. Querying each of these oligonucleotides individually for every genome would require a significant amount of time. To reduce this to an acceptable running time, we group oligonucleotides that have close starting locations on the genome. When we take an oligonucleotide string that cannot be assigned in a group, we assign this oligonucleotide as a key string and give it a group ID, and every other oligonucleotide whose starting point is between the start and end points of that key string is assigned to that group. Later, we randomly selected a user-defined number of sequences for each group. This randomness assures a uniform coverage of amplicons through the genome. In the results section, we report the effects of choosing a different number of representative sequences.

There is also an optional filtering step for oligonucleotides that have a high possibility of self-interaction or interacting with oligonucleotides within the same amplicon. This filtering is performed based on maximum allowed Tm values for homo- or heterodimerization. We also use maximum allowed temperature values where the free energy of interaction becomes zero for pairwise $3'$ end oligonucleotide interactions (Tm where $\Delta G=0$). We find this filtering more reliable than others in vitro. These filtering steps can significantly reduce the number of oligonucleotides and shorten running time. However, it is not recommended to use such a filtering if a large number of potential results are not anticipated, because many unwanted interactions of short oligonucleotides can be avoided by optimizations in the PCR protocol. At the end of this step, every single oligonucleotide is ready to be searched across target and background genomes.

For querying every oligonucleotide on every genome, we construct suffix arrays for each genome using the Mummer4 program and store them. For each individual oligonucleotide, a query is performed against these suffix arrays in different CPU cores.

For each individual oligonucleotide, there may be none or many hits on the queried genomes, and we keep a list of start and end positions of hit regions that we later use 1) to understand how these oligonucleotides form amplicons in these other genomes, 2) to find true interaction strength between these regions and given string. The main reason we first use approximate hit locations of possible oligonucleotides is that finding true interaction strength is computationally expensive as we have mentioned; therefore, we reduce possible true interaction locations before proceeding to the next stage. Finding an approximate location is based on finding a short, exact common string between the oligonucleotide and the genome and extending it based on local alignment. The effects of these parameters are extremely important and are discussed in detail in subsubsection 2.6.1. Subsequently, we use the tool, Primer3 [15], and find interaction properties of every oligonucleotide and its possible hit regions. Using the results of this step, we decide whether an oligonucleotide can be used to discriminate a genome or not. This decision is based on given allowed temperature values and is the same as in extracting oligonucleotides from the seed genome. This thermodynamic interaction analysis step is the most computationally intensive stage and is also parallelized; the operations performed on each genome are executed on separate CPUs.

After finding all oligonucleotides that can hybridize to the target genome in PCR conditions, all amplicons for each genome are calculated and assembled, and we decide whether an oligonucleotide set would amplify given target and non-target genomes. To accept or reject an oligonucleotide set, first, we decide 1) the maximum allowed Tm difference between the interaction of the seed target genome and the oligonucleotide, and the interaction of the input target genome and the oligonucleotide, 2) the minimum allowed Tm difference between the interaction of the seed target genome and the oligonucleotide, and interaction of the input non-target background genome and the oligonucleotide. Finally, the results are filtered according to the required false positive and true positive rates.

2 Results and Evaluation

2.1 Datasets

We applied our method to three highly divergent viruses: 1) Hepatitis C virus (HCV), where the subtypes differ in 31-33% of nucleotide sites on average [16], 2) Human immunodeficiency virus (HIV), which exhibit variations about 25-35% between subtypes and 15-20% within subtypes [17], and 3) the Dengue virus, whose respective genomes (only DENV 1–4) share 60% sequence identity to each other [18]. We chose these viruses because they propose challenges due to their high mutation rate and extensive variability in their genomes.

All complete HIV and HCV genomes have been downloaded from the https://www.hiv.lanl.gov/website. For HIV, entries without sampling year information or with a sampling year prior to 2005 have been removed. Then, genomes with length above 8500bp have been selected. No recombination or subtype filtering has been performed on the sequences. After the filtering, 13838 HIV sequences have been provided as input to the program. For HCV, genomes with lengths above 9300bp have been selected. Also, genomes having non-ACGT bases have been removed. 1657 sequences have been provided as input to the program. Dengue virus genomes are downloaded from the NCBI virus database: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/ with no non-ACGT bases and genome lengths longer than 10500, resulting in 4037 genomes. The reference genome of HIV in the NCBI database is 9181bp, the longest of HCV reference genomes is 9711bp, and the longest of Dengue virus reference genomes is 10735. We chose length limits as 8500, 9300, and 10500 to have close-to-complete input genomes in our dataset.

The raw input genome files along with all result files are provided at the following public GitHub repository: https://github.com/studies-related/Genotype-and-Subtype-Detection-of-Highly-Divergent-Viruses.

2.2 *In Silico* HIV Common Region Study

In our HIV dataset, there are 156 different subtypes and recombinant forms. We aimed to find three oligonucleotides; two primers and one hybridization probe to amplify the highest number of input genomes.

Table 1 shows that the identified oligonucleotides can *in silico* amplify 13801 of 13838 HIV genomes with a true positive rate of 0.997, and this is a remarkable result. When we inspect the false negative results, we see that up to 3 of the remaining 37 genomes could still be amplified depending on the PCR protocol; because their amplicon lengths are about 365-431bp.

The identified amplicon is in a region that codes Integrase and this protein is shown to be the most conserved protein of HIV [19].

2.3 *In Silico* HCV Study

In this experiment, our dataset includes 936 labeled genomes of genotypes 1-6 with further defined subtypes, and 721 genomes are non-labeled. In the first HCV experiment, we aimed at finding a primer and probe set that can amplify genomes maximally. Then in the second part, we attempted to find a primer and probe set that can differentially amplify only given target genotype or subtype while not amplifying any of the other non-target genotypes and subtypes. Table 2 lists the amplicon regions and Table 3 shows their detection/differentiation performances.

The key point here is how the presence or absence of a light signal from three short oligonucleotides in laboratory conditions is defined. In our *in silico* HCV study, we are looking for short regions which have mutated in at least three different subregions compared to other subtypes. This combined approach enhances the robustness and reliability of the analysis.

However, we could not achieve this for HCV subtype 6; the true positive rate was around 80%, and the false positive rate was close to 20%, which is not acceptable. So,

Table 1 HIV Common Region

Genomic location and the sequences	Subtypes	Seed Genome	True Positive Rate
4789-4810—4892-4927—4961-4985 primer1: 5′-AAAAGAAAAGGGGGGATTGGGG-3′ probe: 5′-TTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAG-3′ primer2: 5′-ATTACTACTGCCCCTTCACCTTTCC-3′	156 different subtypes and recombinant forms	Genbank AB287363	13801/13838 0.997

Table 2 Amplicon Regions and Seed Genomes for the HCV Genome

Purpose	Seed Genome	Genomic Location
Differentiate 1a	AB520610	514-534—549-569—722-740
Differentiate 1b	AB016785	9102-9121—9140-9163—9270-9295
Differentiate 2	AB030907	172-196—239-265—274-294
Differentiate 3	AB691595	335-355—555-572—579-597
Differentiate 4	AB795432	178-196—216-230—375-393
Differentiate 5	KF373567	7329-7348—7386-7413—7545-7564
Differentiate 6	D63822	156-175—246-273—277-296
Identify all	AB016785	148-168—285-312—324-343

HCV subtype 6 genomes do not have clear cut short regions that have distinct three mutated subregions conserved among them. However, hybridization probes emit signals from one probe region in PCR; therefore, we thought it is still possible to find a single, highly different region while ignoring primer binding sites. The effect of this change would be, if the amplicons were run on a gel using the old electrophoresis system, it could be observed that they amplify fragments from various other subtypes. Therefore, we lifted restrictions of primers and set the maximum allowed Tm of fluorescent probes binding to non-target genomes to 0 degrees. A probe that binds to this vastly different region would not emit light in PCR under any condition for other subtypes. Thus, we can analyze subtypes based on at least three moderate differences multiplicatively or based on a single significant major difference.

The amplicon that can amplify 0.999% of all HCV genomes *in silico* lies in the 5′UTR region of HCV. These sequences and the predicted secondary structures are highly conserved among HCV genotypes and subtypes [20].

2.4 *In Silico* Dengue Virus Study

We also ran our method on the Dengue Virus, which is another virus that has extensive variation and whose respective genomes (DENV 1-4) share 60% sequence identity to each other [18]. Table 4 shows the seed genomes used and the amplicon regions identifed for the Dengue virus. Table 5 shows the performance of the amplicons.

Table 3 Detection/Differentiation Performance of Amplicons on the HCV Genome

	Hepacivirus C Genotype/Subtype Counts							
	1:727	2:82	3:39	4:20	5:3	6:62	Non-Labeled	
Genotype/ Subtype Differentiation	(1):10 (1a):319 (1b):393 (1c):2 (1g):2 (1l):1	(2):10 (2a):19 (2b):31 (2c):8 (2i):4 (2j):3 (2k):2 (2l):2 (2m):2 (2q):1	(3):2 (3a):32 (3b):2 (3g):1 (3j):1 (3k):1	(4):4 (4a):1 (4g):1 (4k):3 (4l):1 (4m):2 (4q:3) (4r):1 (4s):1 (4v):3	(5a):2 (5):1	(6):19 (6a):17 (6b):2 (6d):1 (6e):2) (6g):1 (6h):1 (6i):1 (6k):1 (6n):2 (6o):1 (6r):2 (6xj):8 (6xi):4	721	True Positive Rate False Positive Rate
Differentiate 1a	1a: 318/319 1\1a: −	−[1]	−	−	−	−		318/319 = 0.997 0/604
Differentiate 1b	1b: 389/393 1\1b: −	−	−	−	−	−		389/393 = 0.990 0/530
Differentiate 2	1b: 2/393 1a: 1/319 1\(1a∪1b): −	2b:30/31 2\2b: +	−	−	−	−		81/82 = 0.988 3/851 = 0.004
Differentiate 3	−	−	+	−	-	−		39/39 0/894
Differentiate 4	−	−	−	+	−	−		20/20 0/913
Differentiate 5	−	−	−	−	+	−		3/3 0/930
Differentiate 6	−	−	−	−	−	+		62/62 0/871
Identify all	+	2b: 30/31 2\2b: +	+	+	+	+	+	1656/1657 = 0.999

[1] − indicates None, + indicates All.

Table 4 Amplicon Regions and Seed Genomes for the Dengue Virus Genome

Purpose	Seed Genome	Genomic Location
Differentiate 1	AB074760	7727-7749—7757-7793—7982-8008
Differentiate 2	AB122020	153-173—182-211—260-286
Differentiate 3	AB189125	2035-2058—2136-2154—2183-2204
Differentiate 4	AF326573	10308-10328—10391-10417—10502-10527
Identify all	AB074760	10477-10496—10508-10531—10587-10614

Table 5 Detection/Differentiation Performance of Amplicons on the Dengue Virus Genome

	Dengue Virus Serotype Counts				
Serotype Differentiation	1:1512	2:1400	3:889	4:215	True Positive Rate False Positive Rate
Differentiate 1	+[1]	−	−	−	1512/1512 0/2507
Differentiate 2	−	1396/1400	−	−	1396/1400 = 0.997 0/2619
Differentiate 3	−	−	887/889	−	887/889 = 0.998 0/3127
Differentiate 4	−	−	−	+	215/215 0/3801
Identify all	1461/1512	1365/1400	789/889	215/215	3833/4016 = 0.954

[1] − indicates None, + indicates All.

For identification of all input genomes, our results show that the most common region that can host three oligonucleotides lies in the 3′UTR region. Alvarez *et al.* show that the 3′ end of the flavivirus genomes folds into a highly conserved stem–loop (3′SL) and detailed analysis of the structure–function of the 3′SL in dengue virus revealed an absolute requirement of this RNA element for viral replication [21]. The 4.6% false negative rate may be both due to the inherent variation among subtypes and partial lack of 3′UTR regions in the input genomes. For differentiation studies, we achieved a minimum of 99.7% true positive rate with 0% false positive rate for all serotypes.

2.5 Application of the Proposed Method for Validation of Data Quality

In our differentiation study of serotype 1, our original dataset contained 1530 input genomes, and our method could not differentiate 18 genomes that came from the very same source. Then, we used the original genotyping tool that the submitters used, Genome Detective [22]. This tool assigns those genomes to serotype 3, so we removed these genomes from the dataset; however, because none of those genomes appeared as false positive in differentiation study of serotype 3, after aligning them we noticed that they are the same sequence submitted 18 times. They are still available in the result file. Likewise, one single genome given as serotype 4 appeared as false positive in the differentiation run of serotype 2 and false negative in the differentiation run of serotype 4. Moreover, two other genomes labeled as serotype 3 appeared as false positives in the differentiation run of serotype 1 and as false negatives in the differentiation run of serotype 3. We again used Genome Detective, and they assigned these genomes to serotype 2 and serotype 1, respectively. We also removed these three genomes that came from the same source.

We also found one result very intriguing. In the differentiation run of serotype 3, our method *in silico* could identify 887 of 889 input genomes. One false negative is the reference genome and the other is the genome that reference genome is constructed upon [23]. However, all subspecies having a common region except their reference genome is extremely strange and we think that the reference genome of Dengue Virus serotype 3 must be reexamined. Because sensitivity and specificity of our method is very high, potential errors in a genome dataset becomes more visible.

We want to emphasize it is extremely important to analyze the quality of data. Like many other programs, our method does not require any human intervention; however, when data quality is questionable, human intervention might be necessary depending on the context and outcome of the study. One other aspect is that human intervention

might be necessary even with high quality data such as in multiplex studies, where studying evolutionary trees beforehand is crucial.

2.6 Searching Oligonucleotides in Genomes

2.6.1 Effects of Sequence Search Parameters

A crucial part of our proposed algorithm is finding hit locations with a suffix array. Simple measures like the edit distance, $3'$ end heuristics, or *k-mer* counting cannot capture the complexity of variation in genomes. The locations and the number of matches or mismatches, deletions, gaps, and combinations of these affect Tm differently; therefore, we must use lenient constraints on sequence similarity.

However, we still need similarity constraints to reduce the number of results; since, Tm and free energy calculation of two non-complement oligonucleotides is computationally expensive. Here, we used Mummer4's *nucmer* command. With that command, we can choose the minimum number of consecutive matches and use those matches as anchors, Mummer4 can perform Smith-Waterman alignment with a specified minimum length.

We investigated the effects of sequence search parameters used with *nucmer*: the minimum length of exact matches and the minimum length of local alignment around those matches, on two small datasets and performed *in silico* differentiation studies. We chose two different sets to show seemingly opposite effects of both variables and required 100% sensitivity and specificity to amplify these effects. The first set is 60 randomly selected Dengue virus serotype 1 genomes as the target genomes and 60 randomly selected Dengue virus serotype 4 genomes as the non-target genomes. The second set comprises 11 genomes of HIV CRF 85_BC against randomly selected 120 HIV genomes. The results and the raw input files are provided on the same GitHub repository given at the end of Subsection 2.1.

In Table 6 and Table 7, we investigate the effect of the minimum length of exact matches and the minimum length of local alignment around those matches on two different datasets. In the upper first part of the tables, we show the effect of anchor length while keeping the minimum length of local alignment constant. In the second half of the tables, we show the effect of the minimum length of local alignment while keeping the anchor length constant. Here, n denotes the length of oligonucleotides. As we have discussed, oligonucleotides are extracted from the seed genome according to constraints, and one such constraint is the length range. Therefore, in a typical study, the length of oligonucleotide length changes 15bp to 40bp, when we can keep this variable as $n - 8$, the minimum length of local alignment changes 7bp to 32bp accordingly.

Table 6 Effects of Sequence Search Parameters in the Differentiation Study of Dengue 1

Minimum length of exact matches (anchor)	Minimum length of local alignment around anchor (n is the length of the oligonucleotide)	Number of amplicons returned as a result	Running Time (in minutes)
5	$n-8$	21503	19
6	$n-8$	9062	12
7	$n-8$	6299	11
8	$n-8$	3576	11
9	$n-8$	1587	12
10	$n-8$	833	24
11	$n-8$	479	24
12	$n-8$	322	24
14	$n-8$	124	23
18	$n-8$	22	23
5	10	32506	653
5	15	30410	53
5	$n-4$	1891	19
5	$n-6$	9644	19
5	$n-8$	21503	19
5	$n-10$	29490	41
5	$n-12$	33848	84

It is obvious that a smaller length conserved region is found more times than a longer length conserved region. So, as the minimum length of the exact match

Table 7 Effects of Sequence Search Parameters in the Differentiation Study of HIV CRF 85_BC

Minimum length of exact matches (anchor)	Minimum length of local alignment around anchor (n is the length of the oligonucleotide)	Number of amplicons returned as a result	Running Time (in minutes)
3	$n-8$	57	106
4	$n-8$	82	31
5	$n-8$	273	19
6	$n-8$	1778	16
7	$n-8$	2662	14
8	$n-8$	4062	13
9	$n-8$	4647	13
10	$n-8$	3999	23
11	$n-8$	2864	22
12	$n-8$	2311	22
14	$n-8$	1942	21
18	$n-8$	1292	21
5	10	54	810
5	15	467	64
5	$n-4$	3297	12
5	$n-6$	1468	15
5	$n-8$	273	19
5	$n-10$	249	30
5	$n-12$	61	55

increases, fewer hit locations are found, which leads to the filtering of possible amplicons in the target genomes phase. As the number of target input genomes increases, and target and non-target genomes differ more, this effect becomes more prominent. We see this clearly in Dengue1 versus Dengue4 study in Table 6. In such studies between divergent genomes, allowing for shorter exact matches mainly promotes sensitivity. However, when the number of target input genomes is small, and target and non-target genomes are phylogenetically closer, as in the HIV CRF 85_BC subtype study, a completely opposite outcome is observed, as shown in Table 7. That is when we increase this minimum match length and find hit locations, many of the found locations that are reported as unique to the target set could still be present in the background genomes. This effect is also present in more divergent genome differentiation, as shown in Table 6. It is possible that found locations may be present in background

genomes. However, it is less important than the differentiation of phylogenetically closer genomes.

We further investigated the effects of minimum length of local alignment around small exact matches. Here Mummer4 only returns results, if minimum length of local alignment equals or exceeds a specified length. We gave predetermined values for this variable and parameterized it based on the oligonucleotide length that is queried. Not giving the alignment length, a predetermined value is important; because primer length may be as small as 10-20 bases while probe lengths can extend to 40-60 bases. A predetermined value would either miss many short oligonucleotides' hit locations or would produce a high number of non-specific regions for long oligonucleotides and increase running time. We used the *-maxmatch-* option, which ensures we use all anchor sequences regardless of their uniqueness in genomes. Also, it is important to use the *-nooptimize-* option; by default, Mummer4 optimizes the alignment, which we find that it does not work well for primer search in our method.

In the Dengue 1 vs Dengue 4 differentiation study in Table 6, we see that as the alignment length requirement increases, the number of amplicons found decreases. It is the same logic with the exact match requirement; although these amplicons are valid, because of alignment length requirements, they are mostly filtered in the target genomes phase. However, in the HIV CRF 85_BC differentiation study in Table 7, we see that as the alignment length requirement increases, the number of found amplicons also increases. In fact, again, more amplicons are formed with smaller alignment requirements; however, this time, because non-target genomes are very close to the target genomes, these amplicons are more and more likely to be found in non-target genomes and filtered in the non-target genome phase. In these types of studies, requiring a small length of general match mainly promotes specificity. Again, this effect is also present in more divergent genome differentiation; as in Table 6, it is possible that found locations may be present in background genomes. However, as input genomes

become divergent and the number of input genomes gets smaller, this effect plays a more important role.

Both effects of both variables are in play, and using a very small length of match similarity is mandatory when differentiating highly divergent viruses. This is why many of the existing methods are not able to handle highly divergent viruses or cannot reach this level of sensitivity and specificity.

We have also added two constant values 10 and 15 instead of parameterized minimum length of local alignment around anchor. We know that as this length decreases, analysis become more reliable. However, in both types of studies when we compare constant 10 and $n - 12$, the results are very close; however, there is a huge difference between running times. This is another benefit of the proposed methology.

2.6.2 Effect of Number of Queried Oligonucleotides per Group

We also investigated how choosing different numbers of oligonucleotides for the queries from every group of oligonucleotides affects the results. For this purpose, we conducted a differentiation study of Dengue 1b subtype (393 genomes) against all Dengue 1 genotypes (530 genomes). The seed genome selected by our method for this study is AB016785. Table 8 shows the amplicon regions and their performances.

In Table 8, we see that randomly choosing a single oligonucleotide (1x) from every group already gives satisfactory results because every part of a genome and all amplicon regions are analyzed. Base differences in that oligonucleotide region make that slight distinction as to whether it can bind to some missed genomes. This effect is clearly visible when we compare 1x and 3x runs. Here, we also see that different runs may output different regions for subtype discrimination. We only reported the best discriminating regions for every run; however, all amplicons above target true positive and false positive rates are reported in the output, and these regions and others are also present in other runs. So, as future work, it could be possible to analyze genomes faster with a slightly lower true positive rate limit and then extract amplicons and

Table 8 Effect of Number of Queried Oligonucleotides Per Group for Differentiating Dengue 1b

Sampling Size	Amplicon Region	True Positive Rate False Positive Rate	Running Time
5x	9050-9072—9104-9124—9275-9294	388/393 = 0.987 0/530	203 minutes
4x	9024-9048—9060-9090—9106-9125	390/393 = 0.992 3/530 = 0.006	157 minutes
3x	437-455—519-547—714-732	386/393 = 0.982 0/530	126 minutes
2x	8974-8996—9040-9068—9108-9127	386/393 = 0.982 0/530	87 minutes
1x	437-455—516-535—714-732	385/393 = 0.979 0/530	51 minutes

surrounding regions from the seed genome and run the analysis again, choosing a larger number of oligonucleotides. We tried this manually, and true positive rates of HIV identification study increased from 99.6% to 99.7%, the HCV identification study increased from 99.8% to 99.9%, and the HCV 1b study increased from 98.7% to 99%. So, although small, this approach increases sensitivity and specificity while significantly reducing running time. As expected, the running time is approximately linear with the number of oligonucleotides.

2.7 Using Common Regions of Reference Genomes

We wanted to assess the performance when our method is used only on common regions. For this purpose, we think the most appropriate inputs are reference genomes. So, from four reference genomes for every serotype of Dengue Virus, NC_002640, NC_001474, NC_001475, and NC_001477, we extracted common regions longer than 15bp in the optional first step of our method, which are shown in Table 9.

Designing three oligonucleotides from these regions can maximally only amplify approximately 37% of all genomes *in silico*. The common region approach loses variation and therefore beneficial information, and the results are significantly below an

Table 9 Common regions of the Dengue Virus

Genomic Location	Sequence
78-94	TAGAGAGCAGATCTCTG
132-149	TCAATATGCTGAAACGCG
10488-10503	GGTTAGAGGAGACCCC
10563-10590	AAGGACTAGAGGTTAGAGGAGACCCCCC
10599-10620	AAACAGCATATTGACGCTGGGA
10622-10643	AGACCAGAGATCCTGCTGTCTC

acceptable threshold. We want to emphasize that, for an important diagnostic work, designing primers only from reference genomes is inefficient because it does not reflect and capture variation. For HCV and HIV, of input genomes there is no single common region greater than 15 base pairs. A motif like representation for reference genomes could be valuable for studies that rely on reference genomes of highly divergent species.

2.8 Experimental Settings

For the experiments, we used a 64-core computer using 60 of them. The HIV study lasted 48 hours and our general parameters, inclusive, are listed in Table 10 below.

Table 10 Parameter ranges used in the experiments

Parameter	Range/Set of Values
primer length range	[19, 30]
primer Tm range	[56, 67]
probe length range	[19, 42]
probe Tm range	[59, 74]
amplicon length range	[80, 350]
anchor length for queries	5
minimum alignment length for queries	max(10, length of oligo-8)
maximum Tm difference between primers	4
minimum Tm difference between primers and probe	-5
minimum primer Tm for target genomes	50
minimum probe Tm for target genomes	55
maximum primer Tm for non-target genomes	45
maximum probe Tm for non-target genomes	50
number of random oligonucleotides chosen from each group for query	4
concentration of monovalent cations	50mM
concentration of divalent cations	3mM
concentration of dNTP	0.8mM
concentration of the primer	800nM
concentration of the probe	400nM
concentration of DNA	50nM

Instead of minimum and maximum Tms for target and non- target genomes, we also implemented minimum/maximum allowed Tm differences between oligonucleotides (primer/probe) and the seed genome, and between oligonucleotides and target/non-target genomes; however, we did not use it. The default values we chose are 10 and 15 degrees difference for primers and probe, respectively for both target and non-target genomes.

2.9 Comparison with Other Studies

We first compared our results to [7]. They tried their method on 2863 Dengue virus genomes. Their method does not take a non-target genome set and instead they use BLAST for assessing specificity; so, we compared the performance on identification of all genomes. Since the dataset they used is not directly available, we were not able to conduct a direct comparison. However, as shown in Subsection 2.4, our true positive rate is 95.4% on 4019 genomes, while the true positive rate of their best performing three oligonucleotide set is 82.3%.

We then compared our method to PrimerHunter [5]. It is a tool specifically designed to differentiate between variable virus subtypes. PrimerHunter could not produce a result on our complete dataset in reasonable time; so, we used a smaller dataset consisting of 50 HCV 1a genomes and 50 HCV 1b genomes. Its run lasted about 18 hours while our method finished in about 30 minutes for these genomes. Our parallel architecture is the main reason behind this running time performance difference. Moreover, PrimerHunter was able to generate 38 different amplicons from 2 non-overlapping regions with maximum true positive rate of 98%, while our method generated 2816 amplicons from 15 different non-overlapping regions, all of them with true positive rates of 100%.

We also compared our method to the method by [11]. For amplification primers, they do not use the machine learning approach proposed in their study, and instead,

they use simple heuristics with mismatch similarity. Their best performing two oligonucleotides to be used as primers achieve 92% accuracy for the HCV dataset, while our method achieves 99.9% using the oligonucleotides given in Table 2.

3 Discussion

3.1 Effects of Transcription Profiles of Viruses

While our method generates oligonucleotides for identifying or subtyping viruses, the discovered oligonucleotides may not reflect the mRNA profile of viruses, and the output would only or mainly depend on the genome. In that case, the amplification signal may be lower than that of primers targeting highly transcribed genes.

We observed this effect on SARS-CoV-2. Because the number of sequenced genomes of Sars-COV-2 have reached millions, we have taken five random genomes from every subtype from the beginning of the pandemic. After analyzing these genomes, it is revealed that the best regions that can, in theory, successfully be used to identify all subtypes are in nsp3, nsp4, nsp5A, nsp13 and nsp16 regions. However, cq -the signal threshold cycle- values of these primers are about 10-15 cq higher than the primers that were published by NIH at the beginning of the outbreak, and they are constructed from the N gene. Although this region is not totally conserved among all subtypes and the true positive rate is lower, the mRNA of this gene is present in most of the open reading frames of the virus, so the amplification signal is higher and cq values are lower. This effect has also been observed and reported in various studies ([24] and [25]). In cases like this, using more than one primer-probe set may be necessary. So, especially for screening studies, knowing inner mechanisms of the target pathogen is highly important.

3.2 Scalability and Future Work

Although our method outputs high resolution results, it cannot yield results within acceptable running times for organisms that are significantly larger than viruses. The most time-consuming stage is the thermodynamic analysis with Primer3. Leber *et al.* developed a fast method for Tm calculation, especially useful for large scale calculation [12]. We believe fast thermodynamic calculation is necessary to target larger genomes. We can argue that there are many successful programs that work better on large genomes; yet a fast Tm calculation can help processing hundreds of thousands of virus genomes or processing tens of thousands of genomes with reasonable running times on personal computers.

We also think that the structure of our method can be improved for subtype analysis. We can combine suffix array query and Tm analysis, and instead of querying every single oligonucleotide to a single target genome in a CPU, we can query one oligonucleotide against every genome. In that case, as soon as the hit count violates the required limits, the program would proceed to another oligonucleotide. For example, if we have a thousand background genomes and our false positive rate limit is 0.005, a sixth hit is enough to eliminate that oligonucleotide. Moreover, since the purpose of subtype analysis is to find a small region that is vastly different and a substantial portion of genome regions are similar, this would reduce running times significantly.

As explained in the results section, it could also be possible to analyze genomes faster with slightly lower true positive rate limit and then extract amplicons and surrounding regions from the seed genome and run the analysis again choosing a larger number of oligonucleotides. Potentially, in addition to reducing the running time, this strategy can improve true positive rates.

As implemented in various tools, generating oligonucleotides from multiple genomes can be beneficial.

Although next-generation sequencing (NGS) technologies are gaining widespread adoption, we posit that polymerase chain reaction (PCR) will remain a critical tool especially in surveillance applications, primarily due to its cost-effectiveness and rapid turnaround time. Moreover, the underlying methodology can be adapted into a subtyping/genotyping tool using NGS data. By identifying all distinctive subregions according to a desired sensitivity and specificity for each subspecies, the classification task can be framed as a probabilistic inference problem. In such a context, a Bayesian classifier — known to minimize the probability of misclassification among classifiers utilizing the same feature set[26] — can be employed effectively, provided that the marginal and conditional probabilities are appropriately estimated.

4 Conclusion

In this study, we presented a methodology that addresses the design of PCR primers and hybridization probes, specifically designed to differentiate specific species, or a set of subspecies from another set of subspecies. What sets this method apart from the existing methods is its unique capability to handle highly divergent viruses. The sensitivity and specificity of our method is also superior to existing state-of-the- art methods. This achievement is made possible through the parallelization of multiple steps and the optimization of intermediate processes. Due to its efficiency, our implementation can process tens of thousands of viral genomes.

The significance of this method extends to various fields that require virus discrimination. These include crucial areas such as public health; screening and tracking viral strains; biomedical research, agriculture, evolutionary studies, and biothreat identification. With some modifications, the methodology can be extended to support oligoarray assays, sequencing studies and can be used as a virus genotyping/ subtyping tool.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The raw input genome files along with all result files are provided at the following public GitHub repository: https://github.com/studies-related/Genotype-and-Subtype-Detection-of-Highly-Divergent-Viruses. The source code of the described methodology is not available as it is owned by the company where the first author worked. However, the first author retains the right to publish the methodology and results presented in this manuscript.

Competing interests

Not applicable

Funding

Not applicable

Authors' contributions

BD conceived and implemented the methodology and ran all the experiments. BD drafted the manuscript. TC supervised the study and edited the manuscript. All authors reviewed the manuscript.

Acknowledgements

Not applicable

References

[1] Li F, Stormo GD. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics. 2001 11;17(11):1067–1076. https://doi.org/10.1093/bioinformatics/17.11.1067.

[2] Rose TM, Henikoff JG, Henikoff S. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Research. 2003 07;31(13):3763–3766. https://doi.org/10.1093/nar/gkg524.

[3] Gadberry MD, Malcomber ST, Doust AN, Kellogg EA. Primaclade—a flexible tool to find conserved PCR primers across multiple species. Bioinformatics. 2004 11;21(7):1263–1264. https://doi.org/10.1093/bioinformatics/bti134.

[4] Jabado OJ, Palacios G, Kapoor V, Hui J, Renwick N, Zhai J, et al. Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. Nucleic Acids Research. 2006 11;34(22):6605–6611. https://doi.org/10.1093/nar/gkl966.

[5] Duitama J, Kumar DM, Hemphill E, Khan M, Măndoiu II, Nelson CE. PrimerHunter: a primer design tool for PCR-based virus subtype identification. Nucleic Acids Research. 2009 03;37(8):2483–2492. https://doi.org/10.1093/nar/gkp073.

[6] Vijaya Satya R, Kumar K, Zavaljevski N, Reifman J. A high-throughput pipeline for the design of real-time PCR signatures. BMC Bioinformatics. 2010 Jun;11(1):340. https://doi.org/10.1186/1471-2105-11-340.

[7] Hysom DA, Naraghi-Arani P, Elsheikh M, Carrillo AC, Williams PL, Gardner SN. Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. PLOS ONE. 2012 04;7(4):1–12. https://doi.org/10.1371/journal.pone.0034560.

[8] Lee HP, Sheu TF. An algorithm of discovering signatures from DNA databases on a computer cluster. BMC Bioinformatics. 2014 Oct;15(1):339. https://doi.org/10.1186/1471-2105-15-339.

[9] Marinier E, Zaheer R, Berry C, Weedmark KA, Domaratzki M, Mabon P, et al. Neptune: a bioinformatics tool for rapid discovery of genomic variation in bacterial populations. Nucleic Acids Research. 2017 08;45(18):e159–e159. https://doi.org/10.1093/nar/gkx702.

[10] Karim S, McNally RR, Nasaruddin AS, DeReeper A, Mauleon RP, Charkowski AO, et al. Development of the Automated Primer Design Workflow Uniqprimer and Diagnostic Primers for the Broad-Host-Range Plant Pathogen Dickeya dianthicola. Plant Disease. 2019;103(11):2893–2902. https://doi.org/10.1094/PDIS-10-18-1819-RE.

[11] Metsky HC, Welch NL, Pillai PP, Haradhvala NJ, Rumker L, Mantena S, et al. Designing sensitive viral diagnostics with machine learning. Nature Biotechnology. 2022 Jul;40(7):1123–1131. https://doi.org/10.1038/s41587-022-01213-5.

[12] Leber M, Kaderali L, Schönhuth A, Schrader R. A fractional programming approach to efficient DNA melting temperature calculation. Bioinformatics. 2005 03;21(10):2375–2382. https://doi.org/10.1093/bioinformatics/bti379.

[13] Onodera K. Selection for 3′-End Triplets for Polymerase Chain Reaction Primers. In: Yuryev A, editor. PCR Primer Design. Totowa, NJ: Humana Press; 2007. p. 61–74.

[14] Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 2018 01;14(1):1–14. https://doi.org/10.1371/journal.pcbi.1005944.

[15] Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Research. 2012 06;40(15):e115–e115. https://doi.org/10.1093/nar/gks596.

[16] Echeverrìa N, Moratorio G, Cristina J, Moreno P. Hepatitis C virus genetic variability and evolution. World J Hepatol. 2015;7(6):831–845. https://doi.org/10.4254/wjh.v7.i6.831.

[17] Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The challenge of HIV-1 subtype diversity. The New England journal of medicine. 2008;358(15):1590–1602. https://doi.org/doi:10.1056/nejmra0706737.

[18] Dwivedi VD, Tripathi IP, Tripathi RC, Bharadwaj S, Mishra SK. Genomics, proteomics and evolution of dengue virus. Briefings in Functional Genomics. 2017 01;16(4):217–227. https://doi.org/10.1093/bfgp/elw040.

[19] Li G, Piampongsant S, Faria NR, Voet A, Pineda-Peña AC, Khouri R, et al. An integrated map of HIV genome-wide variation from a population perspective. Retrovirology. 2015 Feb;12(1):18. https://doi.org/10.1186/s12977-015-0148-6.

[20] Niepmann M, Gerresheim GK. Hepatitis C Virus Translation Regulation. International Journal of Molecular Sciences. 2020;21(7):2328. https://doi.org/10.3390/ijms21072328.

[21] Alvarez DE, De Lella Ezcurra AL, Fucito S, Gamarnik AV. Role of RNA structures present at the 3′UTR of dengue virus on translation, RNA synthesis, and viral replication. Virology. 2005;339(2):200–212. https://doi.org/https:

//doi.org/10.1016/j.virol.2005.06.009.

[22] Vilsker M, Moosa Y, Nooij S, Fonseca V, Ghysens Y, Dumon K, et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. Bioinformatics. 2018 08;35(5):871–873. https://doi.org/10.1093/bioinformatics/bty695.

[23] Peyrefitte CN, Couissinier-Paris P, Mercier-Perennec V, Bessaud M, Martial J, Kenane N, et al. Genetic Characterization of Newly Reintroduced Dengue Virus Type 3 in Martinique (French West Indies). Journal of Clinical Microbiology. 2003;41(11):5195–5198. https://doi.org/10.1128/jcm.41.11.5195-5198.2003.

[24] Rana DR, Pokhrel N, Dulal S. Rational Primer and Probe Construction in PCR-Based Assays for the Efficient Diagnosis of Drifting Variants of SARS-CoV-2. Advances in Virology. 2022;2022(1):1–14. https://doi.org/10.1155/2022/2965666.

[25] Colton H, Ankcorn M, Yavuz M, Tovey L, Cope A, Raza M, et al. Improved sensitivity using a dual target, E and RdRp assay for the diagnosis of SARS-CoV-2 infection: Experience at a large NHS Foundation Trust in the UK. The Journal of infection. 2021;82(1):159–198. https://doi.org/10.1016/j.jinf.2020.05.061.

[26] Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. Springer; 1996.