EFFICIENT PRIMER DESIGN FOR GENOTYPE AND SUBTYPE DETECTION OF HIGHLY DIVERGENT VIRUSES IN LARGE SCALE GENOME DATASETS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

BURAK DEMİRALAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY
IN
HEALTH INFORMATICS

MARCH 2024

EFFICIENT PRIMER DESIGN FOR GENOTYPE AND SUBTYPE DETECTION OF HIGHLY DIVERGENT VIRUSES IN LARGE SCALE GENOME DATASETS

Submitted by BURAK DEMİRALAY in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Health Informatics, Middle East Technical University by,

Prof. Dr. Banu Günel Kılıç ________________
Dean, **Graduate School of Informatics**

Assoc. Prof. Dr. Yeşim Aydın Son ________________
Head of Department, **Health Informatics**

Asst. Prof. Dr. Aybar Can Acar ________________
Supervisor, **Health Informatics, METU**

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son ________________
Health Informatics Dept., METU

Asst. Prof. Dr. Aybar Can Acar ________________
Health Informatics Dept., METU

Prof. Dr. Tolga Can ________________
Computer Science Dept., Colorado School of Mines

Assoc.Prof. Dr. Ercüment Çiçek ________________
Computer Engineering Dept., Bilkent University

Asst. Prof. Dr. Burçak Otlu Sarıtaş ________________
Health Informatics Dept., METU

Date: *11.03.2024*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name :	**Burak Demiralay**
Signature :	**______________**

ABSTRACT

EFFICIENT PRIMER DESIGN FOR GENOTYPE AND SUBTYPE DETECTION OF HIGHLY DIVERGENT VIRUSES IN LARGE SCALE GENOME DATASETS

Demiralay Burak

Ph.D., Department of Medical Informatics

Supervisor: Assist. Prof. Dr. Aybar Can Acar

March 2024, 65 pages

Identification of microorganisms 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 method to extract sequences that would detect the presence of a given species/subspecies using the 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 behind the proposed methodology. We applied our method to three highly divergent viruses: 1) HCV, where the subtypes differ in 31%-33% of nucleotide sites on the average; 2) HIV, for which, 25-35% between-subtype and 15-20% within-subtype variation are observed; and 3) the Dengue virus, whose respective genomes (only DENV 1–4) share 60% sequence identity. Using the proposed method, we were able to select oligonucleotides that can identify 99.9% of 1657 HCV genomes, 99.7% of 11838 HIV genomes, and 95.4% of 4016 Dengue genomes *in silico*. We also show subspecies identification on genotypes 1-6 of HCV and genotypes 1-4 of the Dengue virus with >99.5% true positive and <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 we studied in this thesis.

Keywords: viral diagnostics, genome analysis, primer design, metagenomics

ÖZ

BÜYÜK ÖLÇEKLİ GENOM VERİ KÜMELERİNDE FARKLILAŞMA ORANI YÜKSEK VİRÜSLERİN GENOTİP VE ALTTİP TESPİTİ İÇİN ETKİLİ PRİMER TASARIMI

Demiralay Burak

Doktora, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Yrd. Doç. Dr. Aybar Can Acar

Mart 2024, 65 sayfa

Mikroorganizmaların tanımlanması; tanı çalışmaları, patojen taraması, biyomedikal ve evrimsel çalışmalar, tarımda ve biyolojik tehdit değerlendirmesinde çok önemli bir adımdır. Daha büyük organizmaların incelenmesinde ilerleme kaydedilirken, tek sarmallı virüsler gibi yüksek mutasyon oranlarına ve genetik çeşitliliğe sahip organizmalar için binlerce genomun tamamını işleyebilecek etkili ve ölçeklenebilir bir yönteme ihtiyaç vardır. Bu çalışmada, PCR yöntemini kullanarak belirli bir türün/alt türün varlığını tespit edecek dizilerin çıkarılmasına yönelik bir yöntem geliştirdik. Tüm analizlerde tür tespiti büyük ölçüde ölçüm yöntemine bağlıdır ve PCR'da termodinamik etkileşimler kritik olduğundan, termodinamik önerilen metodolojideki ana itici güçtür. Yöntemimizi kendi içinde oldukça farklılaşmış üç virüse uyguladık; 1) Alt tiplerin nükleotid bölgelerinin ortalama %31-%33 oranında farklılık gösterdiği HCV, 2) Alt tipler arası %25-35 ve alt tip içi varyasyonun %15-20 olduğu HIV ve 3) Genomları (yalnızca DENV 1-4) birbiriyle %60 dizi özdeşliğini paylaşan Dang virüsü. Yöntemimizi kullanarak, 1657 HCV genomunun %99.9'unu, 11838 HIV genomunun %99.7'sini ve 4016 Dang virüsü genomunun %95.4'ünü *in silico* olarak tanımlayabilen oligonükleotidleri seçebildik. Ayrıca ortalama olarak >%99.5 gerçek pozitif ve <%0.05 yanlış pozitif oranıyla HCV'nin 1-6 genotipleri ve Dang virüsünün 1-4 serotipleri üzerinde alt tür tanımlamasını da gösteriyoruz. En gelişmiş yöntemlerin hiçbiri, bu tezde incelediğimiz gibi kendi içinde oldukça farklılaşmış viral genomlar üzerinde bu özgüllük ve hassasiyete sahip oligonükleotidler üretememektedirler.

Anahtar Sözcükler: virüs tanılama, genom analizi, primer tasarımı, metagenomik

To my family

ACKNOWLEDGEMENTS

Without my family; my mother, sister and father, this work wouldn’t be possible.

I also would like to thank Tolga Can for our years long scientific discussions and friendship.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
PCR	Polymerase Chain Reaction
HCV	Hepacivirus C
HIV	Human Immunodeficiency Virus
HPV	Human Papillomavirus
Tm	Melting Temperature
k-mer	Oligonucleotide of length k
CPU	Central Processing Unit

CHAPTER 1

INTRODUCTION

Identification of viruses and bacteria is crucial in various fields. In healthcare and clinical studies, virus and bacteria identification plays a significant role in characterizing pathogens, diagnosing and enabling timely responses to infectious diseases, guiding appropriate treatment strategies, and monitoring disease outbreaks [1]. Identifying viral and bacterial pathogens is necessary to develop new treatments and instrumental in advancing scientific knowledge about infectious diseases. Pathogen identification is also integral to biodefense efforts, enabling the rapid detection of biothreat agents in the event of bioterrorism or biological warfare. Kits developed for this purpose are crucial for surveillance and early warning systems [2]. It is also important in environmental monitoring to detect and identify pathogens in various settings, such as water sources, food production facilities, and agricultural environments to ensure the safety of the environment [3]. Pathogen identification is needed in industrial settings for quality control purposes, particularly in sectors like pharmaceuticals and biotechnology. Identification kits help ensure product safety and compliance with regulatory standards [4]. Overall, accurate and fast virus and bacteria identification is needed in applications in various fields, including healthcare, research, biodefense, environmental monitoring, industrial quality control, and veterinary medicine. It is a must for detection and surveillance across various sectors.

Accurate and timely diagnosis is another critical factor for effective disease management. Traditional diagnostic methods, while valuable, often face limitations in sensitivity, speed, or applicability. In this context, the Polymerase Chain Reaction (PCR) has emerged as a revolutionary technique, transforming the field of diagnostics. PCR boasts remarkable sensitivity, amplifying minute quantities of target DNA or RNA into detectable levels [5]. This surpasses traditional methods like culture-based assays, which can struggle with pathogens that are slow-growing or difficult to cultivate in vitro. Furthermore, PCR delivers results significantly faster, often within hours, than culture methods that may take days or weeks. This rapid turnaround time allows for earlier intervention and improved patient outcomes.

Moreover, PCR paves the way for the development of multiplex assays. These assays can simultaneously detect multiple targets within a single reaction, streamlining the diagnostic process and reducing testing time. This is particularly beneficial when a broad range of potential pathogens must be considered. By overcoming the limitations of traditional methods, PCR has become an indispensable tool in modern diagnostics. Its exceptional sensitivity, speed, and versatility empower healthcare professionals to make informed decisions rapidly, ultimately improving patient care and paving the way for a future of personalized medicine.

Diagnosing viral infections has traditionally posed a challenge due to limitations in sensitivity and differentiation between closely related viruses. This thesis presents a novel approach that significantly enhances our ability to detect and distinguish viruses. This advancement will pave the way for more accurate and timely diagnoses, ultimately improving patient outcomes. Subsequent PCR confirmation can then be employed for definitive identification of the specific viral strain.

Firstly, in the introduction, we will briefly go through necessary biology. Then, in Chapter 2, we look at previous studies in the literature and discuss the shortcomings of previous approaches. In Chapter 3, we describe our proposed method. Chapter 4 is Experimental Results; most of the details of our proposed method are discussed there, and we think it is the most important chapter. Then, in Chapter 5, we discuss future work and conclude. Also in the appendix is the outline of our method’s implementation.

1.1 Background

1.1.1 DNA

Deoxyribonucleic acid (DNA) is the molecule that carries genetic information for the development and functioning of an organism. DNA is made of two linked strands that wind around each other to resemble a twisted ladder — a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four bases: adenine (A), cytosine (C), guanine (G) or thymine (T). The two strands are connected by chemical bonds between the bases: adenine bonds with thymine, and cytosine bonds with guanine. The sequence of the bases along DNA’s backbone encodes biological information [6].

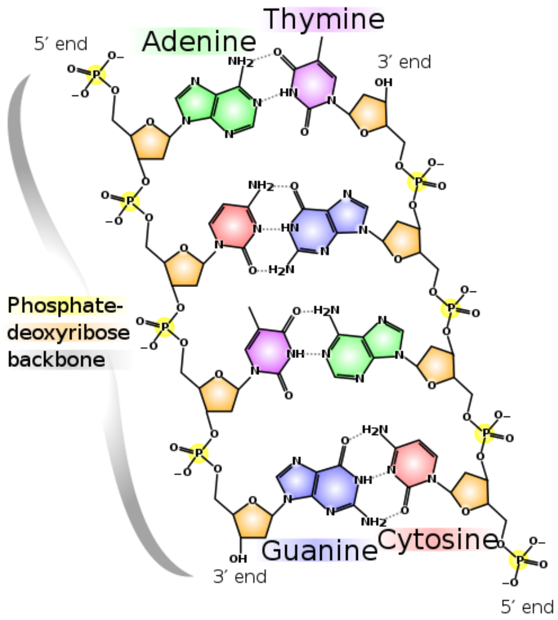


Figure 1: Structure of DNA

As shown in Figure 1, a critical aspect of DNA is that single-stranded DNA string has an asymmetry; two ends called 5'end and 3'end are chemically different, and in PCR, in the presence of a template strand, DNA Polymerase enzyme adds new bases only to the 3'end of the opposing chain.

1.1.2 PCR

Polymerase Chain Reaction (PCR) is a molecular biology technique that amplifies a specific segment of DNA through a series of temperature-dependent enzymatic reactions. This method revolutionized genetic research and diagnostic applications by enabling the rapid and precise replication of DNA sequences.

The PCR process involves three main steps: denaturation, annealing, and extension. Initially, the double-stranded DNA template is heated to a high temperature (typically around 95°C), causing the DNA strands to separate or denature into single strands. This step exposes the target DNA region for subsequent amplification.

Next, the reaction mixture is cooled to a temperature optimal for the binding of short DNA primers to complementary sequences on each strand of the target DNA. This step, known as annealing (typically around 50-60°C), allows the primers to anneal or bind to their specific target sites on the template DNA.

Once the primers are bound, DNA polymerase enzyme, typically Taq polymerase or a thermostable polymerase, extends the primers by adding nucleotides to the 3' end of each primer, synthesizing new DNA strands complementary to the template. This extension step occurs at around 72°C, which is the optimal temperature for DNA polymerase activity.

By repeating these denaturation, annealing, and extension cycles in a specialized thermal cycler machine, the target DNA sequence is exponentially amplified, resulting in a significant increase in the amount of the desired DNA fragment. Each cycle approximately doubles the amount of DNA, allowing for rapid and efficient amplification.

PCR plays a pivotal role in various fields, including medical diagnostics, forensic analysis, evolutionary biology, and genetic engineering. Its unparalleled sensitivity, specificity, and versatility have made PCR an indispensable tool for exploring complex biological processes at the molecular level.

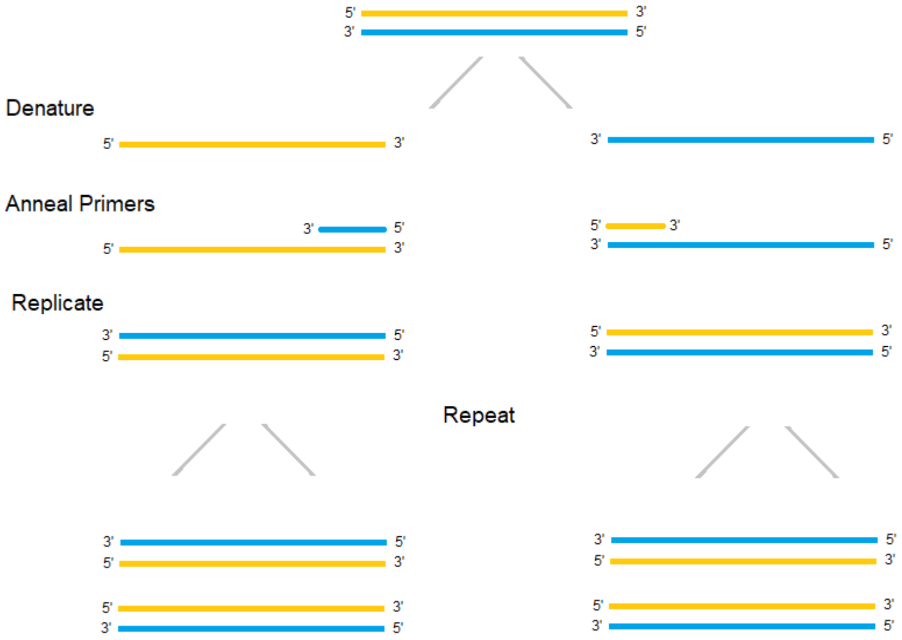


Figure 2: Polymerase Chain Reaction Method

In PCR, there may be a second kind of oligonucleotide other than primers, called probes. Hybridization probes are oligonucleotides designed to attach to the inside of the amplicon region bounded by 3' ends of amplification primers. Only in the presence of amplification they release light.

As shown in Figure 3, TaqMan probes are a type of hybridization probe commonly used in molecular biology and genetics for the specific detection and quantification of nucleic acid sequences in the context of PCR assays. These probes consist of a fluorophore attached to the 5' end and a quencher molecule attached to the 3' end, with a short oligonucleotide sequence complementary to the target DNA or RNA sequence located between them.

During the PCR amplification process, these probes hybridize to the target DNA or RNA template. When the polymerase enzyme encounters the probe-bound target sequence and begins synthesizing new DNA strands, it cleaves the probe, separating the fluorophore from the quencher. This cleavage releases the fluorophore from its proximity to the quencher, resulting in the emission of a fluorescent signal that can be detected in real-time by the PCR instrument and confer quantification of target oligonucleotide reporting light intensity in every cycle.

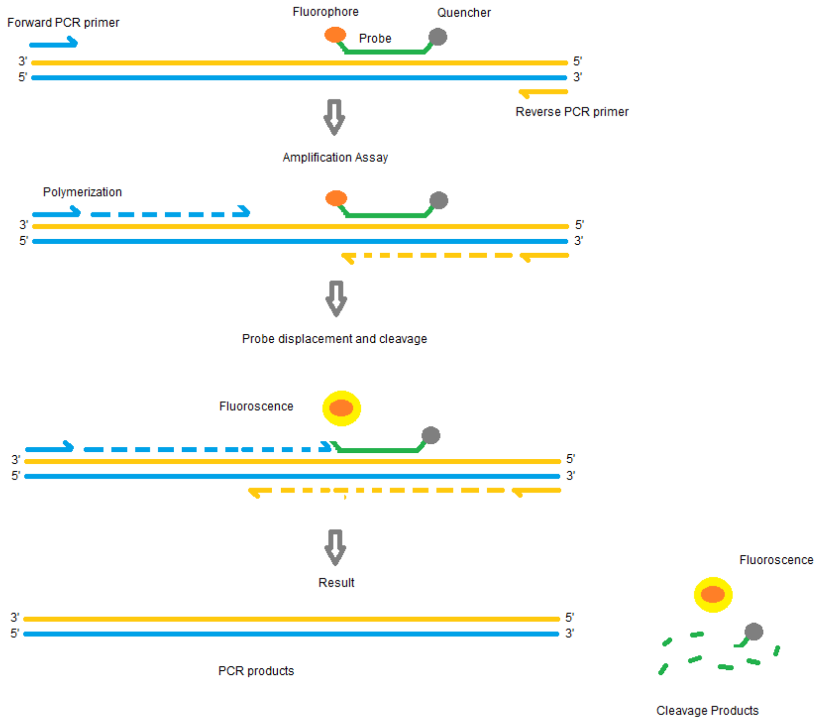


Figure 3: An example of a hybridization probe

1.1.3 Nucleic Acid Thermodynamics

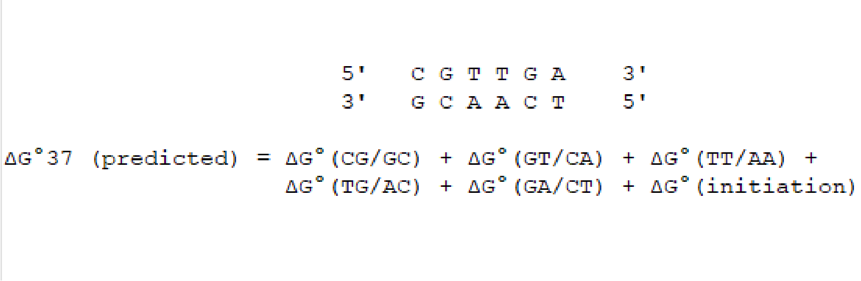
DNA thermodynamics refers to the study of the energetics and stability of DNA molecules and their interactions, particularly focusing on the principles governing the stability of DNA duplexes, hybridization, and nucleic acid secondary structures.

The stability of DNA duplexes, formed by the complementary base pairing of two DNA strands, is influenced by various factors, including temperature, salt concentration, and the sequence and length of the DNA strands. Thermodynamic parameters such as melting temperature (Tm), enthalpy (ΔH), entropy (ΔS), and Gibbs free energy (ΔG) play crucial roles in characterizing the stability of DNA duplexes and predicting their behavior under different experimental conditions. Pioneering work started in the early 1960s as mentioned in the remarkable study by SantaLucia [7].

Understanding DNA thermodynamics is essential for various molecular biology techniques and applications, including polymerase chain reaction and primer design.

For primer/probe interaction with the DNA of the target genome, the melting temperature (Tm) is the most helpful variable. The melting temperature of oligonucleotides, which represents the temperature at which half of the DNA duplexes are denatured into single strands, can be calculated by nearest-neighbor method, which takes into account the sequence composition of the oligonucleotides and considers the contribution of individual base pair interactions to the overall stability of the duplex.

For thermodynamic calculations, base pairs are used and values of parameters are extracted from extensive studies run by different laboratories [7]



```
                        5'   C G T T G A    3'
                        3'   G C A A C T    5'

ΔG°37 (predicted) = ΔG°(CG/GC) + ΔG°(GT/CA) + ΔG°(TT/AA) +
                    ΔG°(TG/AC) + ΔG°(GA/CT) + ΔG°(initiation)
```

Figure 4: An example of free energy calculation

.

Calculation of Tm of an oligonucleotide with a different non-complementary oligonucleotide is far more complex, 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 variables [8]. This calculation is extremely slow; for practical purposes, it is ten thousand times slower than the simple calculation of Tm of an oligonucleotide with its complementary strand. Overall, this calculation is the bottleneck of our proposed method.

Essential practical knowledge that can be derived is that, although not absolutely true and also related to entropy change of interaction, if the oligonucleotide whose Tm of interaction with a target is higher than the same oligonucleotide's Tm of interaction with another target, then the binding affinity of an oligonucleotide with former target is higher at any given temperature.

So, in general, for the PCR method, when we want to design oligonucleotide(s) to amplify one target genome among other similar genomes selectively, we want the Tm of that oligonucleotide with the target to be around or higher than the annealing temperature of PCR, while the Tm of the oligonucleotide with other similar genomes is lower. The critical Tm difference to avoid a false positive result is about 10 °C degrees below the annealing temperature in PCR. However, it depends on the protocol used; annealing time, type of the enzyme, and some reagents have significant effects.

1.2 Problem Statement

A critical challenge in PCR-based diagnostics lies in identifying highly specific primer-probe sets. These sets must fulfill two crucial criteria: 1) Efficient binding and amplification of each target genome within a defined set, and 2) Complete lack of interaction with non-target genomes from a separate background set. This distinction allows for the accurate identification of target organisms.

Our proposed method addresses this challenge by searching for a specific combination of three oligonucleotides. These oligonucleotides consist of one fluorescent probe and two amplification primers. The method seeks to achieve: 1) Target Specificity: The interaction temperature (Tm) between any oligonucleotide and a target genome must fall within a user-defined range suitable for PCR amplification. This ensures efficient amplification even with potential variations and non-complementary sequences within the target genomes. 2) Background Rejection: The Tm of any oligonucleotide interaction with a non-target background genome must be below a minimum threshold. This ensures no amplification of unintended background DNA.

Our method aims to identify highly specific primer-probe sets for accurate target virus detection using PCR by achieving these criteria.

1.3 Contribution of the Thesis

This thesis presents a novel virus subtyping method designed to enhance the accuracy of PCR confirmation studies significantly.

The proposed method leverages a simple yet powerful approach to address limitations faced by existing techniques, particularly their inability to differentiate within highly variable viral landscapes. We will delve into the methodology, showcasing how this simplicity translates into success. Furthermore, we will explore the shortcomings of alternative methods in capturing stable regions within this complex viral environment. This innovative subtyping method represents a significant leap forward in knowledge and paves the way for further optimization and potential applications beyond its current use in PCR confirmation.

CHAPTER 2

LITERATURE REVIEW

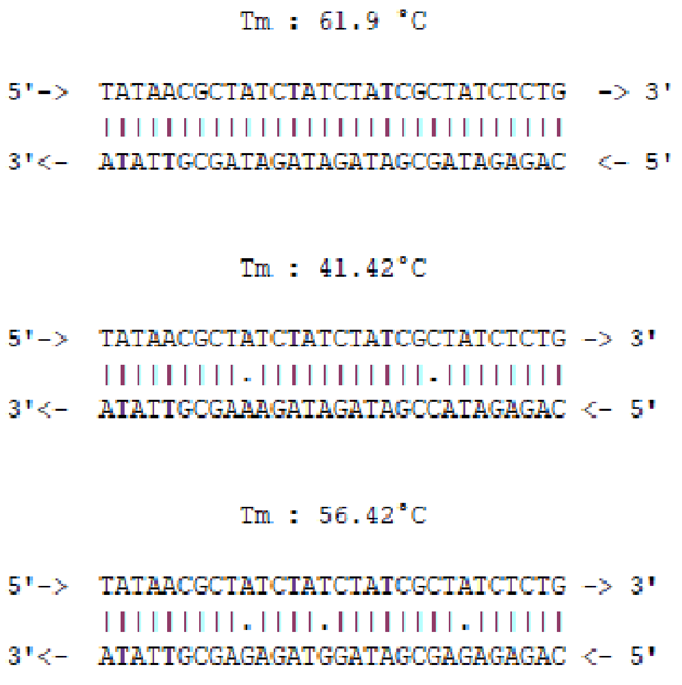
2.1 Previous Studies

DNA signatures are sequences that can distinguish a group of interest from a background group of sequences [9]. 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, evolutionary research, and 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 [10]. Their 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. They calculate the free energy (ΔG) and melting temperatures (Tm) of the potential candidate sequences and select the most discriminating probes based on free energy. In the same year, Rose et al. developed a unique approach called CODEHOP to solve the problem of PCR amplification of distantly related species [11]. They aligned amino acid sequences and employed motif programs, considering codon usage preferences, thereby encoding the conserved amino acid sequences to amplify distantly related species. After finding motifs, they turned amino acid blocks into degenerate primers. In 2004, Gadberry et al. developed the Primaclade tool for identifying conserved PCR primers across multiple species [12]. They used multiple sequence alignment (MSA) of the target sequences, compared individual oligonucleotides to alignment consensus sequences, and scored them according to degeneracy. Because MSA does not work well with a high number of inputs, they suggested preliminary clustering of similar sequences. In 2006, Jabado et al. proposed a method for designing degenerate primers for viruses [13]. To decrease the negative implications of degeneracy, they modeled the problem as a set cover problem and sought 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. They identify the consensus sequences for every branch, score them against all others, and find 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 [14]. They emphasize that degenerate primer approaches ignore primer specificity, which prevents their use in 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 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 [15]. 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 works well on bacteria but not for highly variable viruses. In 2012, Hysom et al. proposed a method that extracts k-length oligonucleotides from all targets and counts them [16]. Their tool picks the most conserved k-mers and realigns them to targets while allowing mismatches. Then, they iteratively find other primer pairs for the remaining targets. In 2014, Lee and Sheu proposed an algorithm and employed a divide-and-conquer strategy and a parallel signature discovery approach [17]. They define a signature, a fixed length l with allowed mismatches d, and this (l,d) pattern, which must occur only once. They recursively divide a given genome into pieces until a full pattern search can run directly on that piece. After finding patterns in each piece, they merge and eliminate the ones found on any other piece. In 2017, Marinier et al. developed Neptune to identify differentially abundant genomic content in bacterial populations [18]. It uses fixed size k-mer matching with a probabilistic model to find the best cardinality of k. They use BLAST for further refinement. In 2019, Karim et al. developed a primer design pipeline, Uniqprimer, to distinguish target genomes from non-target genomes [19]. They first align one target genome to all non-target genomes and extract non-aligned regions. Then, they align these regions with another target region and iteratively align common regions with all target genomes. They design primers from these conserved regions. In 2022, Metsky et al. developed a pipeline for virus amplification, where they use multiple sequence alignment and then, with neural networks, solve a complicated scoring function for the activity of a probe [20].

2.2 Comparison with Previous Approaches

We think interaction analysis to reveal whether two given DNA strands will hybridize must be based solely on thermodynamics because thermodynamics governs interactions of sequences in a laboratory setting. 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 also emphasize that oligonucleotide design based on the number of mismatches may be very misleading. As presented in more detail in the Appendix, a random 25 bp oligonucleotide has a 0.086 probability that its interaction with its complementary sequence with five mutations has higher Tm than its interaction with its complementary sequence with three mutations, and this probability increases to 0.203 when Tm difference is within 5°C which cannot be an accepted difference for differentiating subtypes. In Fig. 5, we show an example of this condition.



```
                Tm : 61.9 °C

5'->   TATAACGCTATCTATCTATCGCTATCTCTG  -> 3'
       ||||||||||||||||||||||||||||||
3'<-   ATATTGCGATAGATAGATAGCGATAGAGAC  <- 5'


                Tm : 41.42°C

5'->   TATAACGCTATCTATCTATCGCTATCTCTG -> 3'
       |||||||||.|||||||||||.||||||||
3'<-   ATATTGCGAAAGATAGATAGCCATAGAGAC <- 5'


                Tm : 56.42°C

5'->   TATAACGCTATCTATCTATCGCTATCTCTG -> 3'
       |||||||||.||||.||||||||.||||||
3'<-   ATATTGCGAGAGATGGATAGCGAGAGAGAC <- 5'
```

Figure 5: Melting temperatures for different number of mismatches

Fig. 5 shows that an oligonucleotide's interaction with its complementary sequence has a much higher binding affinity when there is two vs three mismatches, with a 15°C difference.

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 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 [21].

This heuristic has more potential and 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 last base of the 3’end. However, this probability increases to 0.30 when the Tm difference is kept at 5°C. Therefore, this heuristic also carries the inherent probability of generating false positives in differentiation studies and could reduce true positive rate. We argue that any similarity-based heuristics cannot capture these binding affinities. We have simulated different mutation conditions and show the results in Appendix A, Number of Mutations and Melting Temperature Relation.

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 multiple whole genome sequences 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. Being dependent on sequence similarity heuristics, none of the methods in the literature can efficiently produce oligonucleotides with high specificity and sensitivity on thousands of divergent viral genomes.

CHAPTER 3

PROPOSED METHOD

Our problem involves identifying primer-probe sets that can specifically bind and amplify each genome in a given set of target genomes, T, for PCR, while not binding to any genomes in another set of background genomes, B, thereby identifying the organisms in T. Figure 6 shows the steps of our proposed solution, which are subsequently described in the following subsections.

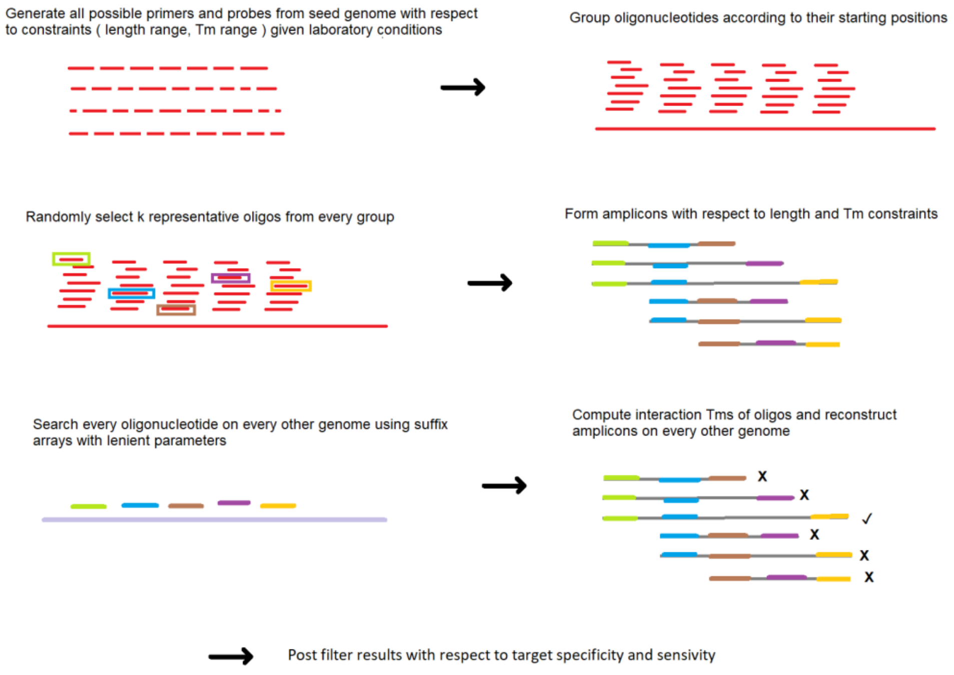

Generate all possible primers and probes from seed genome with respect to constraints (length range, Tm range) given laboratory conditions
Group oligonucleotides according to their starting positions
Randomly select k representative oligos from every group
Form amplicons with respect to length and Tm constraints
Search every oligonucleotide on every other genome using suffix arrays with lenient parameters
Compute interaction Tms of oligos and reconstruct amplicons on every other genome
x
x
✓
x
x
x
Post filter results with respect to target specificity and sensivity

Figure 6: The steps of the proposed method

Pre-filtering

Users can filter the given genomes based on the number of non-ACGT bases they contain or the length of the input. When the number of input genomes is very high, which is the case in many surveillance and diagnostics studies involving various strands of viruses, a tool that can work on them would significantly impact global public health. It is good practice to only choose complete or almost complete genomes for such tasks.

Constructing a Consensus Genome

This is an optional step in which a consensus genome is generated from the common regions of a given subset of the genome set. In this step, common regions are extracted and reassembled using Mummer4 [22]. This process is repeated via stepwise pairwise alignment. The stage of finding common regions among all or a subset of target genomes is primarily added to shorten the running time for organisms with large genomes, such as bacteria, or organisms with low mutation rates. For viruses with high mutation rates, it would be more suitable to include only a single genome for execution so that the consensus genome would be the single provided genome.

The variable in this step is the shortest oligonucleotide length of consensus subregions. As default, we look for at least 15 bp long oligonucleotides. We also convert all non-consensus bases to ‘N’ to avoid missing the location information because amplicon length will be important in next stages.

Extracting Oligonucleotides

This step extracts oligonucleotides that satisfy the given constraints from input genomes. This stage is an oligonucleotide scan in a sliding window fashion. 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, primer, probe, and the DNA.

Generating Amplicons and Choosing Oligonucleotides

We generate the amplicons formed by the oligonucleotides identified in the previous stage. Generally, 100-300 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 yielded 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 with other oligonucleotides within the same amplicon. This filtering is performed based on maximum allowed Tm values for homo- or hetero- dimerization. We also use maximum allowed temperature values where the free energy of interaction becomes zero for pairwise 3’ end oligonucleotide interactions (Tm where ΔG=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 recommended to use this filtering if a large number of potential results are anticipated because many unwanted interactions of short oligonucleotides can be avoided by optimizations in PCR protocol.

At the end of this step, every single oligonucleotide is ready to be searched across target and background genomes.

Querying Oligonucleotides

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, so we reduce possible true interaction locations before going on to the next stage. Finding an approximate location is based on finding a short, exact common string between oligonucleotide and genome and extending it based on local alignment. The effects of these parameters are extremely important and will be discussed in detail in 4.6.1 Effects of Sequence Search Parameters.

Subsequently, we use the tool, Primer3 [23], and find interaction properties of every oligonucleotide and its possible hit regions. This step is extremely important and 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, so the operations performed on each genome are executed on separate CPUs.

Post-filtering

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. Then, the results are filtered according to the required false positive and true positive rates.

The Appendix outlines our method's implementation, focusing on key details.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1. The Dataset

We applied our method to three highly divergent viruses: 1) HCV, where the subtypes differ on average in 31%-33% of nucleotide sites [24]. 2) HIV, which exhibits variations of about 25-35% between subtypes and 15-20% within subtypes [25], and 3) the Dengue virus, whose respective genomes (only DENV 1–4) share 60% sequence identity [26].

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 a size 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 with 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.

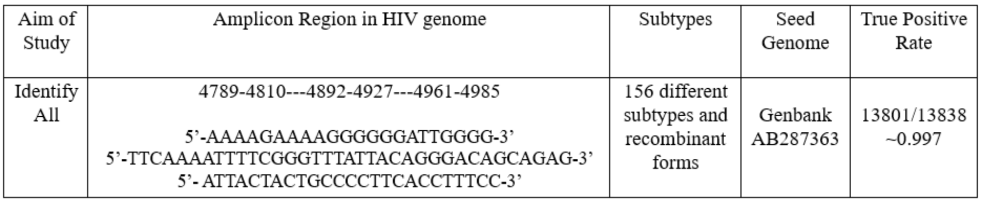
We chose these viruses because they propose challenges due to their high mutation rate and extensive variability in their genomes. The reference genome of HIV in the NCBI database is 9181bp, the longest of HCV reference genomes is 9711bp, and the longest of Dengue virus reference genomes is 10735. We chose length limits of 8500, 9300, and 10500 to have close-to-complete input genomes in our dataset.

The raw input genome files, along with the output result files, are provided in Supplementary Data and at the following link: https://github.com/Burak1Demiralay

4.2. 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 I: Amplicon region and their detection performance on the HIV genome



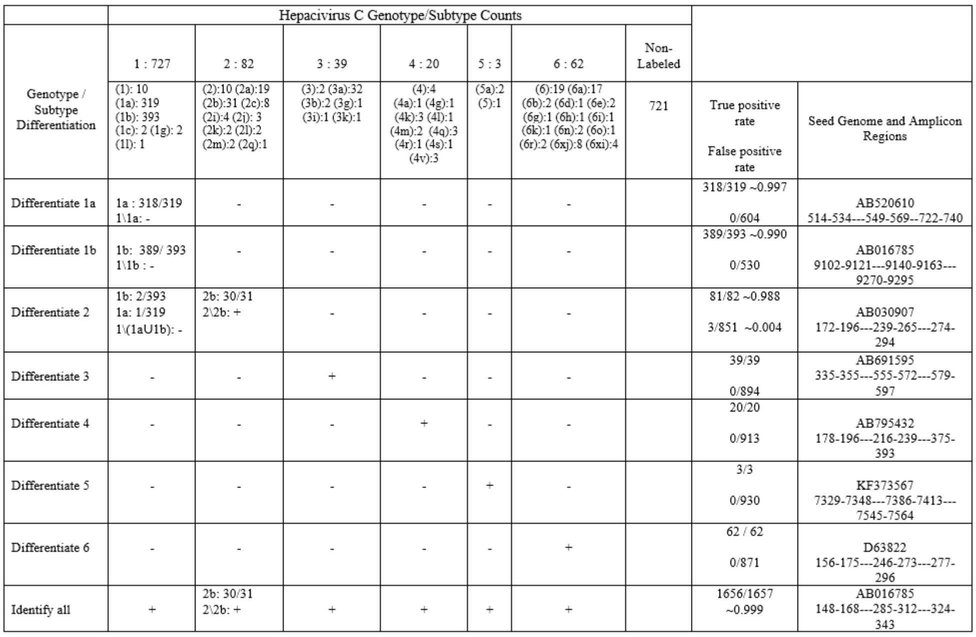
Aim of Study	Amplicon Region in HIV genome	Subtypes	Seed Genome	True Positive Rate
Identify All	4789-4810---4892-4927---4961-4985 5’-AAAAGAAAAGGGGGGATTGGGG-3’ 5’-TTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAG-3’ 5’- ATTACTACTGCCCCTTCACCTTTCC-3’	156 different subtypes and recombinant forms	Genbank AB287363	13801/13838 ~0.997

Table I shows that the identified oligonucleotides can 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. Results are presented with high resolution in the Supplementary Data. The identified amplicon is in a region that codes Integrase, which is shown to be the most conserved protein in HIV [27].

4.3 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 then 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 II lists the amplicon regions and shows their detection/differentiation performance. Using our tool, researchers can also extract unique or defining regions of target species/subspecies for further evolutionary, drug design, or similar studies. We envision the proposed method to be especially useful for less studied organisms.

Table II: Amplicon regions and their detection/differentiation performance on the HCV genome



	Hepacivirus C Genotype/Subtype Counts								
Genotype / Subtype Differentiation	1 : 727	2 : 82	3 : 39	4 : 20	5 : 3	6 : 62	Non-Labeled		
	(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 (3i):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	Seed Genome and Amplicon Regions
Differentiate 1a	1a : 318/319 1\1a: -	-	-	-	-	-		318/319 ~0.997 0/604	AB520610 514-534---549-569--722-740
Differentiate 1b	1b: 389/ 393 1\1b : -	-	-	-	-	-		389/393 ~0.990 0/530	AB016785 9102-9121---9140-9163---9270-9295
Differentiate 2	1b: 2/393 1a: 1/319 1\(1aU1b): -	2b: 30/31 2\2b: +	-	-	-	-		81/82 ~0.988 3/851 ~0.004	AB030907 172-196---239-265---274-294
Differentiate 3	-	-	+	-	-	-		39/39 0/894	AB691595 335-355---555-572---579-597
Differentiate 4	-	-	-	+	-	-		20/20 0/913	AB795432 178-196---216-239---375-393
Differentiate 5	-	-	-	-	+	-		3/3 0/930	KF373567 7329-7348---7386-7413---7545-7564
Differentiate 6	-	-	-	-	-	+		62 / 62 0/871	D63822 156-175---246-273---277-296
Identify all	+	2b: 30/31 2\2b: +	+	+	+	+		1656/1657 ~0.999	AB016785 148-168---285-312---324-343

The non-labeled genomes are only included in the common region study. Similarly, ten ‘genotype1’ genomes with no subtype information are not included in differentiation studies of subtypes 1a and 1b. Detailed results are present in Supplementary Data.

The key point here is how the presence or absence of a light signal from three short oligonucleotides (primer pairs and the probe) in laboratory conditions is defined. In our *in-silico* HCV study, we required that the melting temperature of each primer and probe to non- target genomes be a maximum of 45 and 50 degrees, respectively. That allows us to search for short regions mutated in at least three different subregions compared to other subtypes. Since every subregion is distinctive and defining, the analysis is highly reliable.

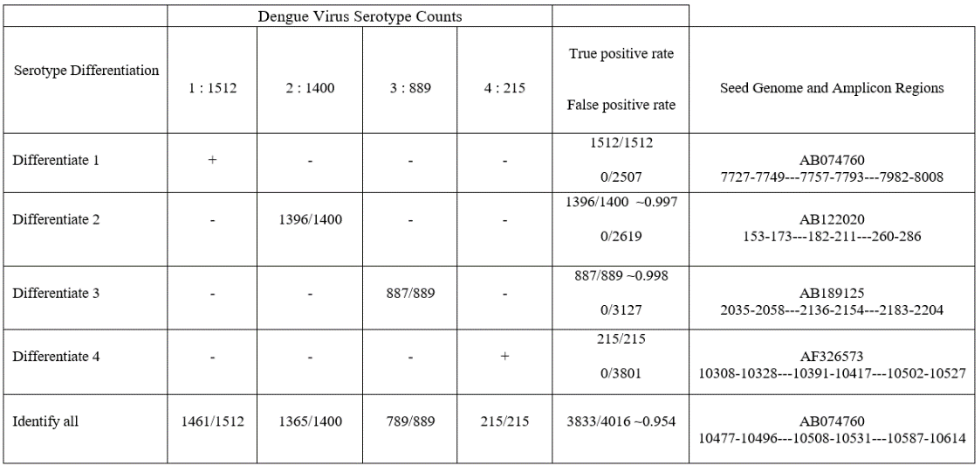
However, we could not achieve this for HCV subtype 6; the true positive rate was around 80%, and the false positive rate was close to 20%, which is not acceptable. So, HCV subtype 6 genomes do not have clear cut short regions that have three distinct mutated subregions conserved among them. However, because hybridization probes emit signals from one region in PCR, 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 5’UTR region of HCV. These sequences and the predicted secondary structures are highly conserved among HCV genotypes and subtypes [28].

4.4 In Silico Dengue Virus Study

We also ran our method on the Dengue Virus, another virus that has extensive variation and whose respective genomes (DENV 1-4) share 60% sequence identity [25]. Table III shows the results of this study.

Table III: Amplicon regions and their detection/differentiation performance 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	Seed Genome and Amplicon Regions
Differentiate 1	+	-	-	-	1512/1512 0/2507	AB074760 7727-7749---7757-7793---7982-8008
Differentiate 2	-	1396/1400	-	-	1396/1400 ~0.997 0/2619	AB122020 153-173---182-211---260-286
Differentiate 3	-	-	887/889	-	887/889 ~0.998 0/3127	AB189125 2035-2058---2136-2154---2183-2204
Differentiate 4	-	-	-	+	215/215 0/3801	AF326573 10308-10328---10391-10417---10502-10527
Identify all	1461/1512	1365/1400	789/889	215/215	3833/4016 ~0.954	AB074760 10477-10496---10508-10531---10587-10614

For the 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* shows 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 [29]. We did not further inspect the percentage of 4.6% false negative rate due to genetic variation, low quality sequencing, or even complete or partial lack of 3’UTR region in input genomes. For differentiation studies, we achieved a minimum of 99.7% true positive rate with 0% false positive rate for all serotypes.

4.5 Application of the Proposed Method for Validation of Data Quality

In the differentiation study of serotype 1 of Dengue Virus, our original dataset contained 1530 input genomes, and our method could not differentiate 18 genomes from the same source. Then, we looked deeper and used the original genotyping tool the submitters used, Genome Detective [30]. That tool assigns those genomes to serotype 3, so we removed these genomes from the dataset; however, because none of those genomes appeared as false positives in the differentiation study of serotype 3, we aligned them, and it appears that they are the same sequence submitted 18 times. They are still available as a result of Supplementary Data. Likewise, one single genome given as serotype 4 appeared as a false positive in the differentiation run of serotype 2 and a false negative in the differentiation run of serotype 4.

Moreover, two other genomes given as serotype 3 appeared as false positives in the differentiation run of serotype 1 and 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. This shows that because the sensitivity and specificity of our method are very high, potential errors in a genome dataset become more visible. In other instances of false negatives or positives, when either the Genome Detective tool could not assign or there was a conflict between their assignment and our analysis, we did not take any action.

We found one result very intriguing. In the differentiation run of serotype 3, the proposed *in silico* method could identify 887 of 889 input genomes. One false negative is the reference genome, and the other is the genome the reference genome is constructed upon [31]. 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.

We want to emphasize that analyzing data and understanding the cause of the observed variation is extremely important. 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. Human intervention may be necessary even with high-quality data. We will discuss this in section 5.2, Effect of Transcription Profile of Viruses, and in Appendix, HPV In Vitro Study Results.

4.6 Searching Oligonucleotides on Genomes

4.6.1 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 using 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. Results are presented in Supplementary Data and the raw input files are provided.

Table IV: 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 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

Table V: 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

In Table IV and Table V, 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.

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 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 IV. 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 V. 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 actually still be present in the background genomes. This effect is also present in more divergent genome differentiation, as shown in Table IV. 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 also 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 10s of bases while probe lengths can extend to 40s of 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 IV, 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 V, 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 IV, 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.

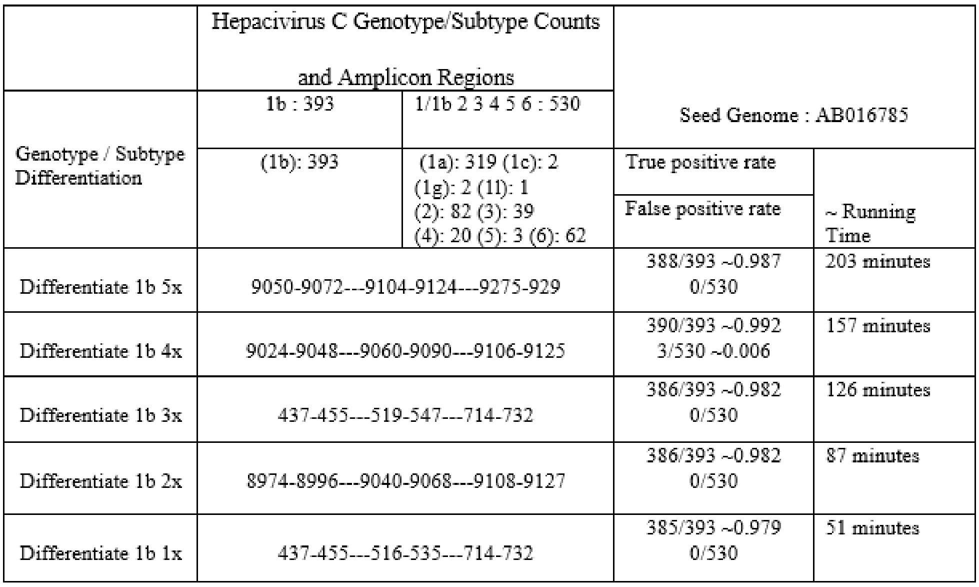
Both effects of both variables are in play, and using a very small length of match similarity is mandatory when differentiating highly divergent viruses. This is why many of the existing methods are not able to handle highly divergent viruses or cannot reach this level of sensitivity and specificity.

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, results are very close, however there is huge difference between running times. This is another innovation and contribution to this field.

4.6.2 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.

Table VI: Effect of number of queried oligonucleotides per group



	Hepacivirus C Genotype/Subtype Counts and Amplicon Regions		Seed Genome : AB016785	
Genotype / Subtype Differentiation	1b : 393	1/1b 2 3 4 5 6 : 530		
	(1b): 393	(1a): 319 (1c): 2 (1g): 2 (1l): 1 (2): 82 (3): 39 (4): 20 (5): 3 (6): 62	True positive rate False positive rate	~ Running Time
Differentiate 1b 5x	9050-9072---9104-9124---9275-929		388/393 ~0.987 0/530	203 minutes
Differentiate 1b 4x	9024-9048---9060-9090---9106-9125		390/393 ~0.992 3/530 ~0.006	157 minutes
Differentiate 1b 3x	437-455---519-547---714-732		386/393 ~0.982 0/530	126 minutes
Differentiate 1b 2x	8974-8996---9040-9068---9108-9127		386/393 ~0.982 0/530	87 minutes
Differentiate 1b 1x	437-455---516-535---714-732		385/393 ~0.979 0/530	51 minutes

In Table VI, we see that randomly choosing a single oligonucleotide from every group (1x) 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, combining these two facts as future work, it could be possible to analyze genomes faster with a 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. 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, it increases sensitivity and specificity while significantly reducing running time. As expected, running time is almost linear with the number of oligonucleotides.

4.7. 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.

78-94 TAGAGAGCAGATCTCTG

132-149 TCAATATGCTGAAACGCG

10488-10503 GGTTAGAGGAGACCCC

10563-10590 AAGGACTAGAGGTTAGAGGAGACCCCCC

10599-10620 AAACAGCATATTGACGCTGGGA

10622-10643 AGACCAGAGATCCTGCTGTCTC

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 acceptable threshold. We also 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 >15bp. A motif like representation for reference genomes could be valuable for studies that rely on reference genomes of highly divergent species.

4.8. Experimental Settings

For the experiments, we used a 64-core computer using 60 of them, so it would be appropriate to limit the input genome count to one-tenth for a similar study to be completed in similar durations on a regular home computer. The HIV study lasted 48 hours (about 2 days) and our general parameters, inclusive, are as follows;

- 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 every group to be queried: 4
- Concentrations of monovalent cations, divalent cations, dNTP, primer, probe, and the DNA: 50mM, 3mM, 0.8mM, 800nM, 400nM, 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.

4.9. Comparison With Other Studies

We first compared our results to Hysom et al. [15]. 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 the Experimental Results section, our true positive rate is 95.4% on 4016 genomes, while the true positive rate of their best performing three oligonucleotide set is 82.3%.

We then compared our method to PrimerHunter[13]. It is a tool specifically designed to differentiate between variable virus subtypes. It 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%. Results are presented with high resolution in Supplementary Data.

We also compared our method to the method by Metsky et al. [19]. 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 three oligonucleotides given in the Experimental Results section.

4.10. In Vitro Validation

We carried out a HPV subtyping study in collaboration with a commercial firm. We are requested to find primers to detect and differentiate HPV subtypes 16, 18, 45, 31, 33, 35, 39, 51, 52, 56, 58, 59, 68, 69 in multiplex PCR where probes to detect different subtypes are tagged with different fluorescent dyes, so it is possible to detect a specific subtype in the presence of other oligonucleotides.

Because subtypes 16, 18 and 45 are more prevalent [32], oligonucleotides to find these three subtypes are expected to run in one PCR well together while others in second well and it is sufficient to confirm presence of any one of second subtype set without differentiating any further.

In our first try, we found out that there is no region in HPV that would be present in genomes of these given subtypes while not present in all other HPV subtypes that we used as background genomes, so there is no single oligonucleotide set that can detect given HPV subtypes without generating false positive results. Therefore, we used the phylogenetic tree to cluster different subtypes and repeated the clustering many times on different branches until the subtypes uniquely shared a common amplicon region that is not present in other background subtype genomes whether among studied or not. So as studies like this, human intervention is necessary not to generate false-positive results.

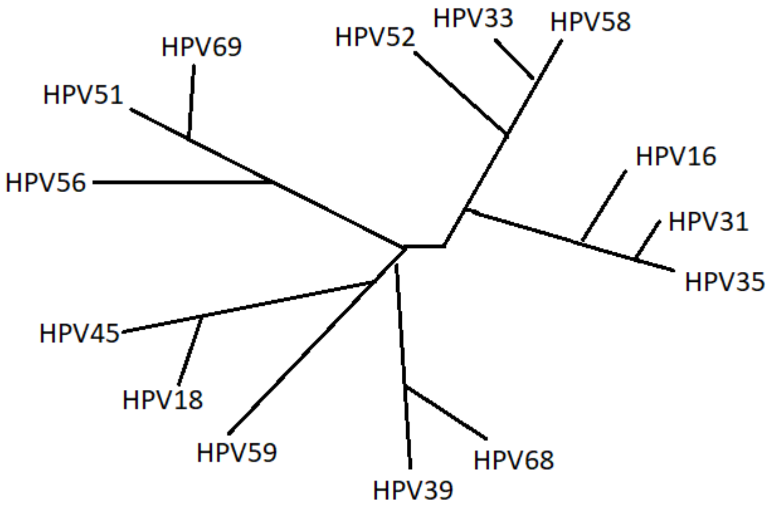


Figure 7: Phylogenetic tree of studied Human Papillomavirus subtypes [33]

We found different oligonucleotides set from three regions to uniquely identify HPV-16 (regions E1, E7, L1), two regions for HPV-18 (regions E7, L1) and two regions for HPV-45 (regions E2, E6) and oligonucleotides for all other subtypes are extracted from E1 region. We designed one common set for detecting subtypes 33-52-58, one set for 31-35, one set for 51-69, one set for 39-59-68 and one for subtype 56.

Positive control mixtures containing subtypes 16, 18 and 45 are obtained from Triplex International Biosciences (Fujian, China) and Microbiologics (Saint Cloud, Minnesota). The PCR amplification was performed in a 20-µl volume containing 4x SensiFAST™ Probe No-ROX Kit, 500nM concentration of each primer and 100nM probe. Amplification and detection were performed by using the Biorad CFX 384 detection system (Applied Biosystems). The amplification ramp included an initial hold step of 10 min at 95°C, followed by a two-step cycle consisting of 15 s at 95°C and 1 min at 57°C, a total of 35 cycles.

At the writing of this thesis, preliminary experiments were just performed and this section was added at the very end of the thesis writing deadline. Therefore, we did not perform a thorough validation including copy number analysis. We think preliminary results are very promising. Here we show one result of one positive control containing subtypes 16, 18, and 45. All other results are shown in the Appendix section. As a follow up study analysis of swap samples from patients are performed, where selected primer-probe sets were successful to distinguish among subspecies (unpublished data through personal communication).

For this study, we designed two tubes, one containing four primer-probe sets differentially targeting and detecting subtypes HPV-16, HPV-18, HPV-45, and all of HPV 39-59-68. The probe of each set was tagged with a different dye emitting in a non-overlapping wavelength range. The second tube contains primer-probe sets differentially targeting and detecting subtypes all of HPV 31-35, all of HPV 33-52-58, all of 51-56-69, and internal control RNase P gene. Because it was not possible to add HPV-56 to any existing cluster without generating false positives, we designed a unique oligonucleotide set for this subtype. However, an existing dye was used for its probe.

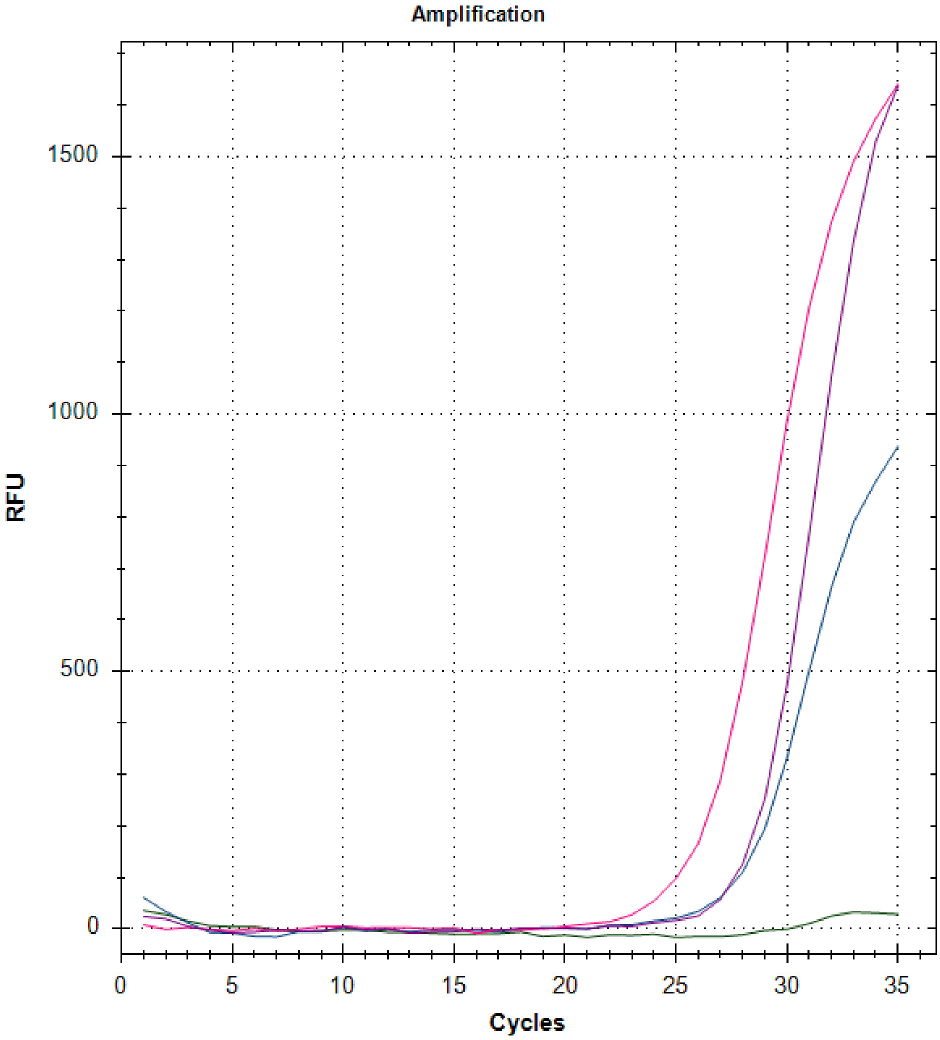


Figure 8: Positive Control Sample, oligonucleotide sets of Tube 1

Figure 8 shows the output of multiplex study of Tube 1 on a positive control sample which contains mixture of subtypes HPV-16, HPV-18 and HPV-45. Every oligonucleotide set differentially amplifies its corresponding target except oligonucleotide set of HPV-39-59-68.

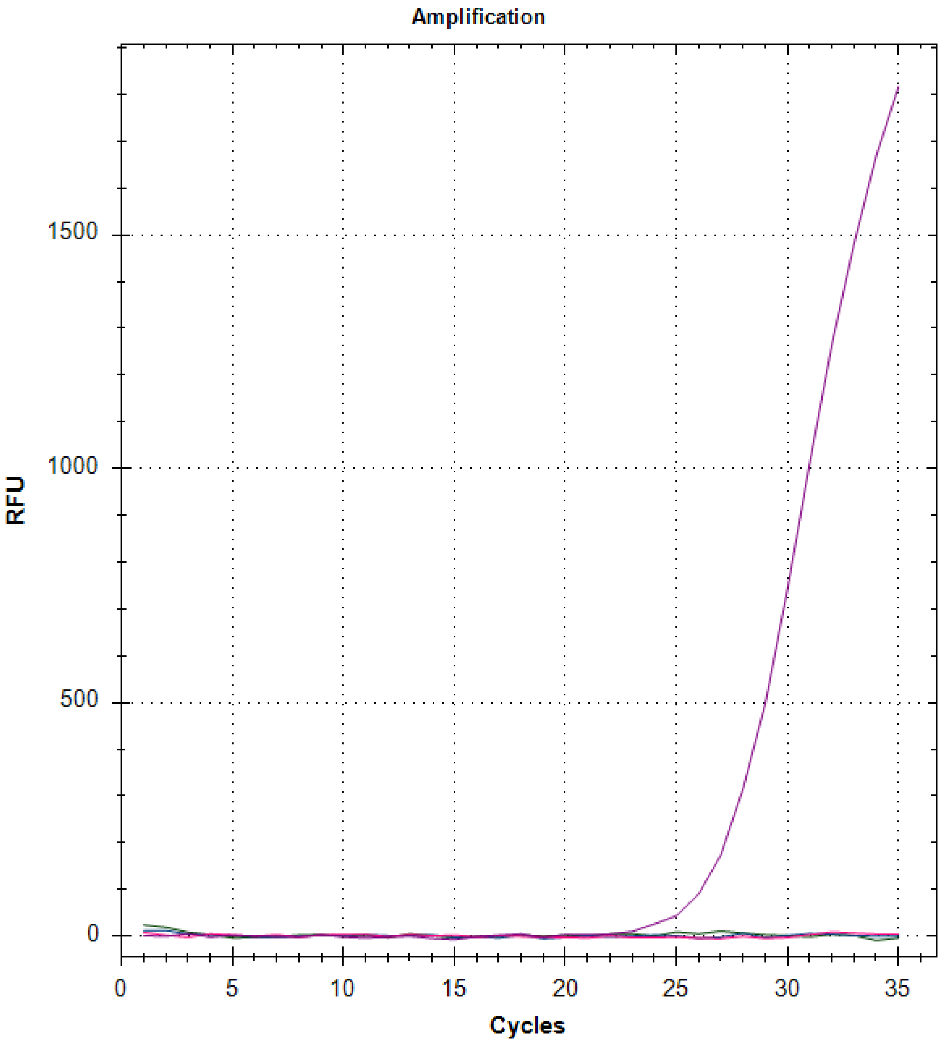


Figure 9: Positive Control Sample, oligonucleotide sets of Tube 2

Figure 9 shows the output of the multiplex study of Tube 2 on a positive control sample, which contains a mixture of subtypes HPV-16, HPV-18, and HPV-45. The amplified signal is from the internal control gene Human RNase P. Control genes are used to check the validity of PCR protocol and sample collection.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 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. We have used Mummer4[22] for string matching, and it is an efficient program that constructs a suffix array with time and space linear in the length of the genome. To identify a match, it uses a sublinear approximate string matching algorithm [34] in time proportional to the length of the query sequence and then utilizes a banded Smith-Waterman alignment algorithm [35], which computes the score matrix along a small band around the diagonal to extend match region with time complexity proportional to longer string. After finding a possible interaction region between the queried oligonucleotide and the target genome, we use Primer3[23] to find the exact melting temperature. As we briefly mentioned in the Nucleic Acid Thermodynamics section, extensive calculation is required.

However, all calculations are the same for viruses or any other organism, and the main effecting difference is genome length and, therefore, the number of possible oligonucleotides that can be extracted from the genome. The average size of RNA viruses is about 10 kb [36], while a typical bacterial genome is around 5 million bp [37]. So, an in-silico experiment involving a bacterium, on average, would take 500 times more time than an in-silico experiment involving an RNA virus. And every added unique genome increases time additively. Even if we utilize all CPU cores efficiently for all genomes, we can only decrease this time increase in proportion to the number of cores. So, for huge data sets involving organisms much larger than viruses, this very detailed analysis becomes infeasible.

Because thermodynamic analysis with Primer3 is very slow, Leber et al. developed a fast method for Tm calculation, which is especially useful for large-scale calculation [3]. We believe fast thermodynamic calculation is necessary to step into larger genomes. We can argue that there are many successful programs that work better on large genomes, yet a fast Tm calculation can help process hundreds of thousands of virus genomes or process 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 a 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 the various tools, generating oligonucleotides from multiple genomes can be beneficial.

Finally, our methodology can be a basis for a novel and very powerful genotyping/subtyping tool that incorporates information from all genomes. Now, for the PCR method, we extract three distinctive subregions in close proximity. However, for that purpose, we can extract every distinctive subregion higher than the desired sensitivity and specificity for every subspecies. Then, given an input genome and predefined oligonucleotides for every subspecies, this problem will turn into a simple classification problem.

5.2 Effect of Transcription Profile of Viruses

While our method generates oligonucleotides for identifying or subtyping viruses, found oligonucleotides may not reflect the mRNA profile of viruses, and 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 sequenced genomes of Sars-COV-2 have reached millions, we have taken five random genomes from every subtype from the beginning of the pandemic outbreak (1955 subtypes). 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 primers that were published by NIH at the beginning of the outbreak, and they are constructed from the N gene. Although in time, this region is not totally conserved among all subtypes and the true positive rate is lower, mRNA of this gene is present in most open reading frames of the virus, so the amplification signal is very high and cq values are low. This effect has also been observed and reported in various studies. [38,39] In cases like this, using more than one primer-probe set may be necessary.

So, for screening studies, knowing inner mechanisms of target pathogen is highly important.

5.3 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 are 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 oligo array assays and be used as a virus genotyping/subtyping tool.

REFERENCES

[1] Wang, H., Zhang, W., & Tang, Y.-W. (2022). Clinical microbiology in detection and identification of emerging microbial pathogens: past, present and future. In Emerging Microbes &amp; Infections (Vol. 11, Issue 1, pp. 2579–2589). Informa UK Limited. https://doi.org/10.1080/22221751.2022.2125345

[2] Perlin, D. (2008). Rapid Detection of Bioterrorism Pathogens. In Beyond Anthrax (pp. 317–334). Humana Press. https://doi.org/10.1007/978-1-59745-326-4_16

[3] Nehra, M., Kumar, V., Kumar, R., Dilbaghi, N., & Kumar, S. (2022). Current Scenario of Pathogen Detection Techniques in Agro-Food Sector. In Biosensors (Vol. 12, Issue 7, p. 489). MDPI AG. https://doi.org/10.3390/bios12070489

[4] Karanam, V. R., Reddy, H. P., Subba Raju, B. V., Rao, J. C., Kavikishore, P. B., & Vijayalakshmi, M. (2008). Detection of indicator pathogens from pharmaceutical finished products and raw materials using multiplex PCR and comparison with conventional microbiological methods. In Journal of Industrial Microbiology &amp; Biotechnology (Vol. 35, Issue 9, pp. 1007–1018). Oxford University Press (OUP). https://doi.org/10.1007/s10295-008-0376-z

[5] Garibyan, L., & Avashia, N. (2013). Polymerase Chain Reaction. In Journal of Investigative Dermatology (Vol. 133, Issue 3, pp. 1–4). Elsevier BV. https://doi.org/10.1038/jid.2013.1

[6] https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid

[7] SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. In Proceedings of the National Academy of Sciences (Vol. 95, Issue 4, pp. 1460–1465). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.95.4.1460

[8] M. Leber, L. Kaderali, A. Schonhuth, and R. Schrader, “A fractional programming approach to efficient DNA melting temperature calculation,” Bioinformatics, vol. 21, no. 10, pp. 2375–2382, 2005. doi:10.1093/bioinformatics/bti379.

[9] Phillippy, A. M., Mason, J. A., Ayanbule, K., Sommer, D. D., Taviani, E., Huq, A., Colwell, R. R., Knight, I. T., & Salzberg, S. L. (2007). Comprehensive DNA Signature Discovery and Validation. In A. Rzhetsky (Ed.), PLoS Computational Biology (Vol. 3, Issue 5, p. e98). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pcbi.0030098

[10] F. Li and G.D. Stormo, “Selection of optimal DNA oligos for gene expression arrays,” Bioinformatics, vol. 17, no. 11, pp. 1067–1076, 2001, doi:10.1093/bioinformatics/17.11.1067.

[11] T. M. Rose, J. G. Henikoff, and S. Henikoff, “CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design,” Nucleic acids research, vol. 31, no. 13, pp. 3763–3766, 2003. doi:10.1093/nar/gkg524.

[12] M. D. Gadberry, S. T. Malcomber, A. N. Doust, and E. A. Kellogg, “Primaclade--a flexible tool to find conserved PCR primers across multiple species,” Bioinformatics, vol. 21, no. 7, pp. 1263–1264, 2005. doi:10.1093/bioinformatics/bti134.

[13] O. J. Jabado et al., “Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments,” Nucleic acids research, vol. 34, no. 22, pp. 6605–6611, 2006. doi:10.1093/nar/gkl966.

[14] J. Duitama, D. M. Kumar, E. Hemphill, M. Khan, I. I. Mandoiu, and C. E. Nelson, “PrimerHunter: a primer design tool for PCR-based virus subtype identification,” Nucleic acids research, vol. 37, no. 8, pp/ 2483–2492, 2009. doi:10.1093/nar/gkp073.

[15] R, Vijaya Satya, K. Kumar, N. Zavaljevski, and J. Reifman, “A high-throughput pipeline for the design of real-time PCR signatures,” BMC Bioinformatics, vol. 11, no. 1, 2010. doi:10.1186/1471-2105-11-340.

[16] D. A. Hysom, P. Naraghi-Arani, M. Elsheikh, A. C. Carrillo, P. L. Williams, and S. N. Gardner, “Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments,” PloS one, vol. 7, no. 4, pp. e34560, 2012. doi:10.1371/journal.pone.003456.

[17] H. P. Lee, T.-F. Sheu, “An algorithm of discovering signatures from DNA databases on a computer cluster,” BMC Bioinformatics, vol. 15, no. 1, 2014. doi:10.1186/1471-2105-15-339.

[18] E. Marinier et al., “Neptune: a bioinformatics tool for rapid discovery of genomic variation in bacterial populations,” Nucleic acids research, vol. 45, no. 18, pp. e159, 2017. doi:10.1093/nar/gkx702.

[19] S. Karim et al., “Development of the Automated Primer Design Workflow Uniqprimer and Diagnostic Primers for the Broad-Host-Range Plant Pathogen Dickeya dianthicola,” Plant Disease, vol. 103, no. 11, pp. 2893–2902, 2019. doi:10.1094/pdis-10-18-1819-re.

[20] H. C. Metsky et al., “Designing sensitive viral diagnostics with machine learning,” Nature Biotechnology, vol. 40, no. 7, pp. 1123–1131, 2022. doi:10.1038/s41587-022-01213-5.

[21] Onodera, K. (2007). Selection for 3′-End Triplets for Polymerase Chain Reaction Primers. In PCR Primer Design (pp. 61–74). Humana Press. https://doi.org/10.1007/978-1-59745-528-2_3

[22] G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin, “MUMmer4: A fast and versatile genome alignment system,” PLOS Computational Biology, vol. 14, no. 1, pp. e1005944, 2018. doi:10.1371/journal.pcbi.1005944.

[23] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen, “Primer3--new capabilities and interfaces,” Nucleic acids research, vol. 40, no. 15, pp. e115, 2012. doi:10.1093/nar/gks596.

[24] N. Echeverría, “Hepatitis C virus genetic variability and evolution,” World Journal of Hepatology, vol. 7, no. 6, pp. 831, 2015. doi:10.4254/wjh.v7.i6.831.

[25] B. S. Taylor, M. E. Sobieszczyk, F. E. McCutchan, and S. M. Hammer, “The Challenge of HIV-1 Subtype Diversity,” New England Journal of Medicine, vol. 358, no. 15, pp. 1590–1602, 2008. doi:10.1056/nejmra0706737.

[26] V. D. Dwivedi, I. P. Tripathi, R. C. Tripathi, S. Bharadwaj, S., and S. K. Mishra, “Genomics, proteomics and evolution of dengue virus,” Briefings in Functional Genomics, pp. elw040, 2017. doi:10.1093/bfgp/elw040.

[27] G. Li et al., “An integrated map of HIV genome-wide variation from a population perspective,” Retrovirology, vol. 12, pp. 18, 2015. doi:10.1186/s12977-015-0148-6.

[28] Niepmann, M., & Gerresheim, G. K. (2020). Hepatitis C Virus Translation Regulation. In International Journal of Molecular Sciences (Vol. 21, Issue 7, p. 2328). MDPI AG. https://doi.org/10.3390/ijms21072328

[29] Alvarez, D. E., De Lella Ezcurra, A. L., Fucito, S., & Gamarnik, A. V. (2005). Role of RNA structures present at the 3′UTR of dengue virus on translation, RNA synthesis, and viral replication. In Virology (Vol. 339, Issue 2, pp. 200–212). Elsevier BV. https://doi.org/10.1016/j.virol.2005.06.009

[30] Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., Pauwels, R., Alcantara, L. C., Vanden Eynden, E., Vandamme, A.-M., Deforche, K., & de Oliveira, T. (2018). Genome Detective: an automated system for virus identification from high-throughput sequencing data. In I. Birol (Ed.), Bioinformatics (Vol. 35, Issue 5, pp. 871–873). Oxford University Press (OUP). https://doi.org/10.1093/bioinformatics/bty695

[31] https://www.ncbi.nlm.nih.gov/nuccore/NC_001475

[32] Clifford, G. M., Smith, J. S., Plummer, M., Muñoz, N., & Franceschi, S. (2003). Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. In British Journal of Cancer (Vol. 88, Issue 1, pp. 63–73). Springer Science and Business Media LLC. https://doi.org/10.1038/sj.bjc.6600688

[33] Van Doorslaer, K. (2013). Evolution of the Papillomaviridae. In Virology (Vol. 445, Issues 1–2, pp. 11–20). Elsevier BV. https://doi.org/10.1016/j.virol.2013.05.012

[34] Chang, W.I., Lawler, E.L. Sublinear approximate string matching and biological applications. Algorithmica 12, 327–344 (1994). https://doi.org/10.1007/BF01185431

[35] Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. In Journal of Molecular Biology (Vol. 147, Issue 1, pp. 195–197). Elsevier BV. https://doi.org/10.1016/0022-2836(81)90087-5

[36] Saberi, A., Gulyaeva, A. A., Brubacher, J. L., Newmark, P. A., & Gorbalenya, A. E. (2018). A planarian nidovirus expands the limits of RNA genome size. In S. Perlman (Ed.), PLOS Pathogens (Vol. 14, Issue 11, p. e1007314). Public Library of Science (PLoS). https://doi.org/10.1371/journal.ppat.1007314

[37] Land, M., Hauser, L., Jun, SR. et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15, 141–161 (2015). https://doi.org/10.1007/s10142-015-0433-4

[38] Rana, D. R., Pokhrel, N., & Dulal, S. (2022). Rational Primer and Probe Construction in PCR-Based Assays for the Efficient Diagnosis of Drifting Variants of SARS-CoV-2. In M. Jabir (Ed.), Advances in Virology (Vol. 2022, pp. 1–14). Hindawi Limited. https://doi.org/10.1155/2022/2965666

[39] Colton, H., Ankcorn, M., Yavuz, M., Tovey, L., Cope, A., Raza, M., Keeley, A. J., State, A., Poller, B., Parker, M., de Silva, T. I., & Evans, C. (2021). 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. In Journal of Infection (Vol. 82, Issue 1, pp. 159–198). Elsevier BV. https://doi.org/10.1016/j.jinf.2020.05.061

APPENDICES

APPENDIX A

Number of Mutations and Melting Temperature Relation

Here we randomly generated a random 25bp oligonucleotide and introduced different number random mutations and calculated Tm's of interaction of original oligonucleotide with mutated oligonucleotides. For every mutation number pairs we generated 50000 oligonucleotides. We simulated variability in genomes; in general we would expect smaller number of mutations accumulated in a region in target genomes than in same region in background genomes, however in subtype differentiation studies of phylogenetically close highly variable genomes, this distinction becomes blurred.

In Table VII, we report the probability when Tm of oligonucleotide with more mutations is higher. For example an oligonucleotide has 0.053 probability that its interaction with its complementary sequence with 7 mutations has higher Tm than its interaction with its complementary sequence with 4 mutations.

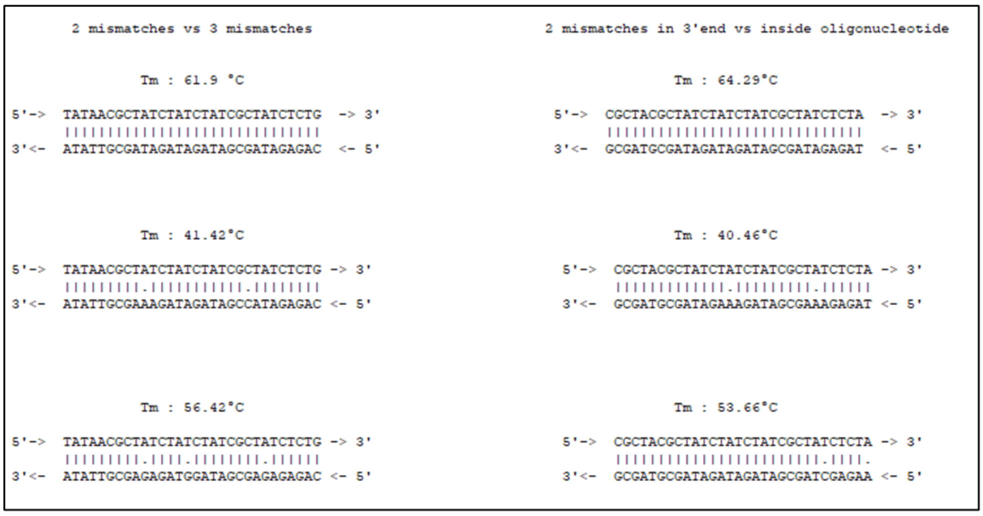
In Table VIII, we report the ratio when Tm of oligonucleotide with more mutations is within 5°C. For example an oligonucleotide has 0.106 probability that its interaction with its complementary sequence with 7 mutations has Tm that is within 5°C of its interaction with its complementary sequence with 4 mutations.

Table VII: Probability of fewer mutations having higher Tm

# of mutations # of mutations	2	3	4	5	6	7	8	9	10
1	0.127	0.037	0.012	0.005	0.002				
2		0.182	0.063	0.023	0.011	0.006			
3			0.22	0.086	0.038	0.018	0.01		
4				0.251	0.114	0.053	0.026	0.016	
5					0.28	0.142	0.073	0.038	0.023

Table VIII: Probability of fewer mutations having Tm within 5°C

# of mutations # of mutations	2	3	4	5	6	7	8	9	10
1	0.521	0.156	0.044	0.015	0.005				
2		0.469	0.184	0.063	0.024	0.011			
3			0.454	0.203	0.084	0.036	0.017		
4				0.449	0.225	0.106	0.049	0.027	
5					0.449	0.249	0.132	0.068	0.038



```
      2 mismatches vs 3 mismatches                              2 mismatches in 3'end vs inside oligonucleotide

                Tm : 61.9 °C                                              Tm : 64.29°C

5'->  TATAACGCTATCTATCTATCGCTATCTCTG  -> 3'              5'->  CGCTACGCTATCTATCTATCGCTATCTCTA  -> 3'
      ||||||||||||||||||||||||||||||                           ||||||||||||||||||||||||||||||
3'<-  ATATTGCGATAGATAGATAGCGATAGAGAC  <- 5'              3'<-  GCGATGCGATAGATAGATAGCGATAGAGAT  <- 5'


                Tm : 41.42°C                                              Tm : 40.46°C

5'->  TATAACGCTATCTATCTATCGCTATCTCTG -> 3'                5'->  CGCTACGCTATCTATCTATCGCTATCTCTA -> 3'
      |||||||||.|||||||||||.||||||||                            |||||||||||||.|||||||||.||||||
3'<-  ATATTGCGAAAGATAGATAGCCATAGAGAC <- 5'                3'<-  GCGATGCGATAGAAAGATAGCGAAAGAGAT <- 5'


                Tm : 56.42°C                                              Tm : 53.66°C

5'->  TATAACGCTATCTATCTATCGCTATCTCTG -> 3'                5'->  CGCTACGCTATCTATCTATCGCTATCTCTA -> 3'
      |||||||||.||||.||||||||.||||||                            ||||||||||||||||||||||||.||||.
3'<-  ATATTGCGAGAGATGGATAGCGAGAGAGAC <- 5'                3'<-  GCGATGCGATAGATAGATAGCGATCGAGAA <- 5'
```

Figure 10: Melting temperatures for different mismatch conditions

Figure 10 shows two cases, left side of figure is an example of mismatch disparity that we have shown in simulation that an oligonucleotide with higher number of mutations can bind to its target much more efficiently.

Right side of figure shows an example that, an oligonucleotide having mutations at 3' end of an oligonucleotide, mutated even at the very tip of oligonucleotide, can bind to its target much more efficiently compared to an oligonucleotide having mutations further from 3' end.

Therefore, all sequence similarity based methods are inherently prone to generating false positive and false negative results.

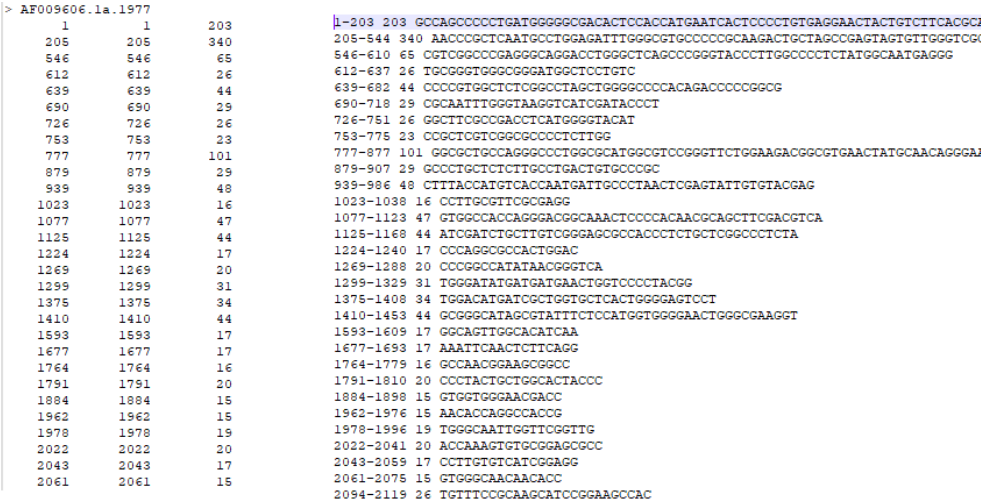
APPENDIX B

Detailed Outline of the Implementation of the Proposed Method

Implementation of our method is an automated pipeline with no human intervention from beginning to end. We designed it in parts so after finishing, one could restart from any middle step. However here we focus on some key points that we think either algorithmically or practically will be useful to anyone who wishes to rewrite and improve our method.

We will step by step go over the implementation on the same dataset that we used in the comparison study; 50 HCV 1a genomes as target genomes and 50 HCV 1b genomes as non-target genomes, however with tighter Tm ranges. Also, here we use three seed genomes only to show generation of consensus genome.

After pre-filtering, we want to construct a consensus genome. It is an iterative process that with Mummer4, we align the result with the new genome in a pairwise manner. We find common regions of length >15bp with the mummer command: # mummer -maxmatch -l 15 -n save location+id ref_file qry_file # save command will generate files as id.aux id.isa id.kmer id.lcp id.sa, after the first run, instead of ‘save’ we use ‘load’ with the same id.

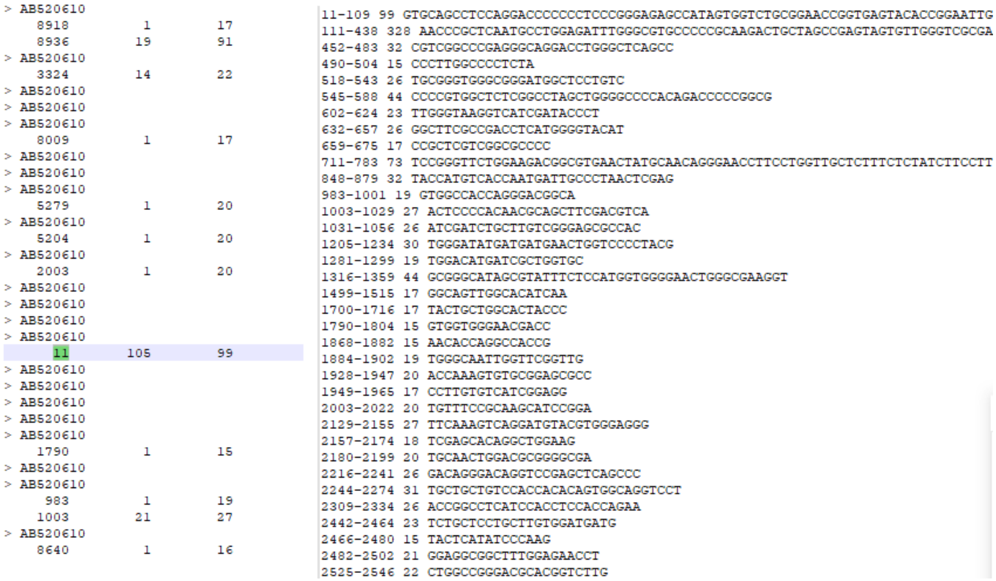


```
> AF009606.1a.1977
       1          1        203
     205        205        340
     546        546         65
     612        612         26
     639        639         44
     690        690         29
     726        726         26
     753        753         23
     777        777        101
     879        879         29
     939        939         48
    1023       1023         16
    1077       1077         47
    1125       1125         44
    1224       1224         17
    1269       1269         20
    1299       1299         31
    1375       1375         34
    1410       1410         44
    1593       1593         17
    1677       1677         17
    1764       1764         16
    1791       1791         20
    1884       1884         15
    1962       1962         15
    1978       1978         19
    2022       2022         20
    2043       2043         17
    2061       2061         15
```

```
1-203 203 GCCAGCCCCCTGATGGGGGCGACACTCCACCATGAATCACTCCCCTGTGAGGAACTACTGTCTTCACGC
205-544 340 AACCCGCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCAAGACTGCTAGCCGAGTAGTGTTGGGTCG
546-610 65 CGTCGGCCCGAGGGCAGGACCTGGGCTCAGCCCGGGTACCCTTGGCCCCTCTATGGCAATGAGGG
612-637 26 TGCGGGTGGGCGGGATGGCTCCTGTC
639-682 44 CCCCGTGGCTCTCGGCCTAGCTGGGGCCCCACAGACCCCCGGCG
690-718 29 CGCAATTTGGGTAAGGTCATCGATACCCT
726-751 26 GGCTTCGCCGACCTCATGGGGTACAT
753-775 23 CCGCTCGTCGGCGCCCCTCTTGG
777-877 101 GGCGCTGCCAGGGCCCTGGCGCATGGCGTCCGGGTTCTGGAAGACGGCGTGAACTATGCAACAGGGA
879-907 29 GCCCTGCTCTCTTGCCTGACTGTGCCCGC
939-986 48 CTTTACCATGTCACCAATGATTGCCCTAACTCGAGTATTGTGTACGAG
1023-1038 16 CCTTGCGTTCGCGAGG
1077-1123 47 GTGGCCACCAGGGACGGCAAACTCCCCACAACGCAGCTTCGACGTCA
1125-1168 44 ATCGATCTGCTTGTCGGGAGCGCCACCCTCTGCTCGGCCCTCTA
1224-1240 17 CCCAGGCGCCACTGGAC
1269-1288 20 CCCGGCCATATAACGGGTCA
1299-1329 31 TGGGATATGATGATGAACTGGTCCCCTACGG
1375-1408 34 TGGACATGATCGCTGGTGCTCACTGGGGAGTCCT
1410-1453 44 GCGGGCATAGCGTATTTCTCCATGGTGGGGAACTGGGCGAAGGT
1593-1609 17 GGCAGTTGGCACATCAA
1677-1693 17 AAATTCAACTCTTCAGG
1764-1779 16 GCCAACGGAAGCGGCC
1791-1810 20 CCCTACTGCTGGCACTACCC
1884-1898 15 GTGGTGGGAACGACC
1962-1976 15 AACACCAGGCCACCG
1978-1996 19 TGGGCAATTGGTTCGGTTG
2022-2041 20 ACCAAAGTGTGCGGAGCGCC
2043-2059 17 CCTTGTGTCATCGGAGG
2061-2075 15 GTGGGCAACAACACC
2094-2119 26 TGTTTCCGCAAGCATCCGGAAGCCAC
```

Figure 11: Common regions of first two genomes

After some scripting, we extract all common subregions and locations. If there are more than two seed genomes, later it is important not to batch query all strings because we noticed that Mummer4 has a bug that may miss already present strings, but this bug is present only in the batch command. So, we query all distinct strings separately, then combine them.



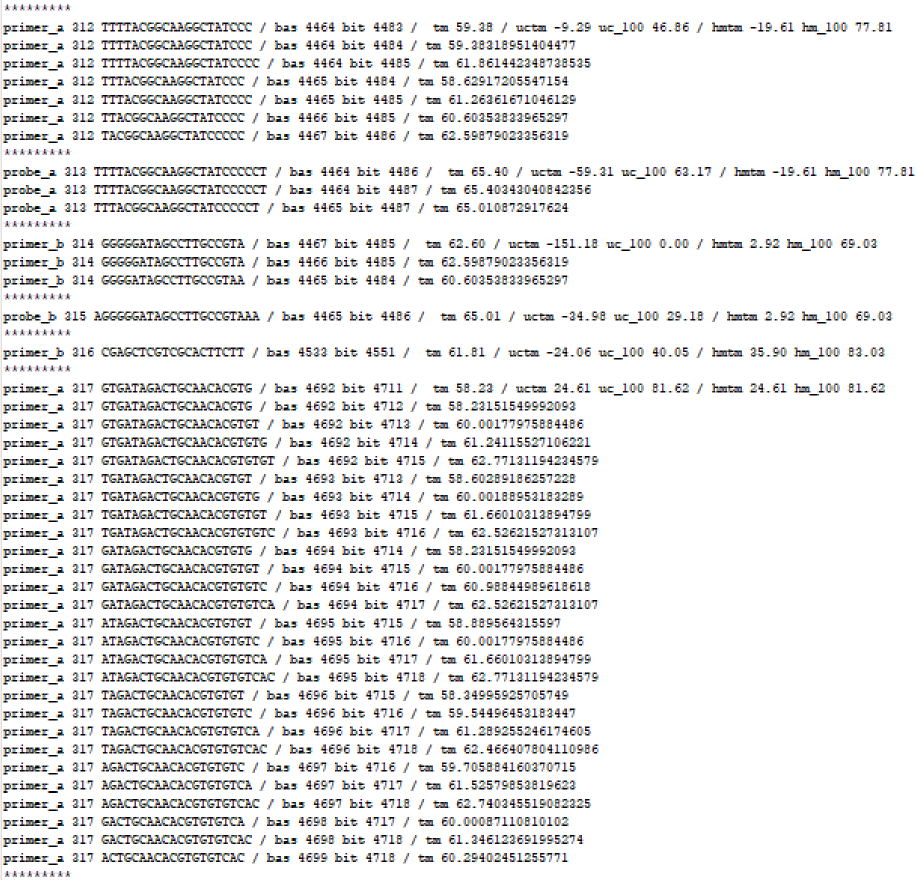
```
> AB520610
    8918       1       17
    8936      19       91
> AB520610
    3324      14       22
> AB520610
> AB520610
> AB520610
    8009       1       17
> AB520610
> AB520610
> AB520610
    5279       1       20
> AB520610
    5204       1       20
> AB520610
    2003       1       20
> AB520610
> AB520610
> AB520610
> AB520610
      11     105       99
> AB520610
> AB520610
> AB520610
> AB520610
> AB520610
    1790       1       15
> AB520610
> AB520610
     983       1       19
    1003      21       27
> AB520610
    8640       1       16
```

```
11-109 99 GTGCAGCCTCCAGGACCCCCCCTCCCGGGAGAGCCATAGTGGTCTGCGGAACCGGTGAGTACACCGGAATTG
111-438 328 AACCCGCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCAAGACTGCTAGCCGAGTAGTGTTGGGTCGCGA
452-483 32 CGTCGGCCCGAGGGCAGGACCTGGGCTCAGCC
490-504 15 CCCTTGGCCCCTCTA
518-543 26 TGCGGGTGGGCGGGATGGCTCCTGTC
545-588 44 CCCCGTGGCTCTCGGCCTAGCTGGGGCCCCACAGACCCCCGGCG
602-624 23 TTGGGTAAGGTCATCGATACCCT
632-657 26 GGCTTCGCCGACCTCATGGGGTACAT
659-675 17 CCGCTCGTCGGCGCCCC
711-783 73 TCCGGGTTCTGGAAGACGGCGTGAACTATGCAACAGGGAACCTTCCTGGTTGCTCTTTCTCTATCTTCCTT
848-879 32 TACCATGTCACCAATGATTGCCCTAACTCGAG
983-1001 19 GTGGCCACCAGGGACGGCA
1003-1029 27 ACTCCCCACAACGCAGCTTCGACGTCA
1031-1056 26 ATCGATCTGCTTGTCGGGAGCGCCAC
1205-1234 30 TGGGATATGATGATGAACTGGTCCCCTACG
1281-1299 19 TGGACATGATCGCTGGTGC
1316-1359 44 GCGGGCATAGCGTATTTCTCCATGGTGGGGAACTGGGCGAAGGT
1499-1515 17 GGCAGTTGGCACATCAA
1700-1716 17 TACTGCTGGCACTACCC
1790-1804 15 GTGGTGGGAACGACC
1868-1882 15 AACACCAGGCCACCG
1884-1902 19 TGGGCAATTGGTTCGGTTG
1928-1947 20 ACCAAAGTGTGCGGAGCGCC
1949-1965 17 CCTTGTGTCATCGGAGG
2003-2022 20 TGTTTCCGCAAGCATCCGGA
2129-2155 27 TTCAAAGTCAGGATGTACGTGGGAGGG
2157-2174 18 TCGAGCACAGGCTGGAAG
2180-2199 20 TGCAACTGGACGCGGGGCGA
2216-2241 26 GACAGGGACAGGTCCGAGCTCAGCCC
2244-2274 31 TGCTGCTGTCCACCACACAGTGGCAGGTCCT
2309-2334 26 ACCGGCCTCATCCACCTCCACCAGAA
2442-2464 23 TCTGCTCCTGCTTGTGGATGATG
2466-2480 15 TACTCATATCCCAAG
2482-2502 21 GGAGGCGGCTTTGGAGAACCT
2525-2546 22 CTGGCCGGGACGCACGGTCTTG
```

Figure 12: Common regions of first three genomes

As we mentioned previously, this common region forming step is optional and if one seed genome is used, that whole genome becomes the consensus string that oligonucleotides will be extracted from. Not constructing consensus genome must be default choice for highly variable viruses because constructing consensus genome may and will lose very important information.

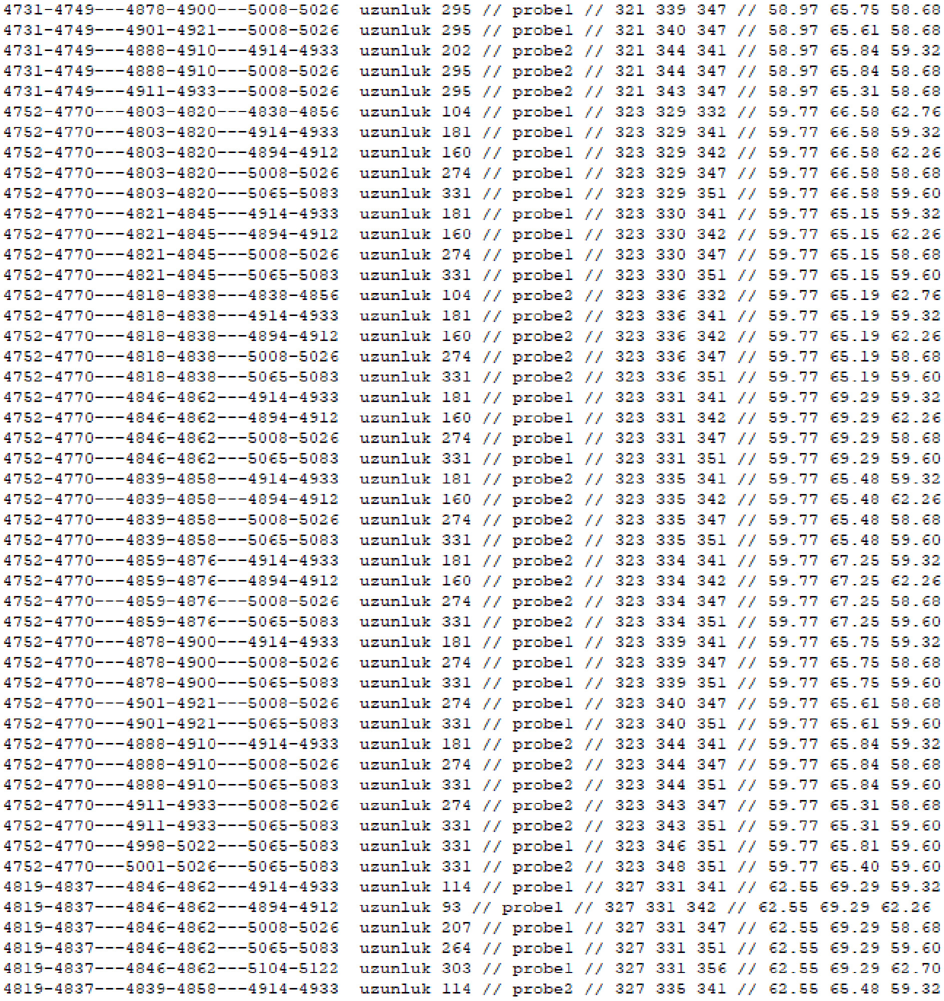
Then we extract oligonucleotides from the common consensus region. It is simply a two-way string scan. We report all primers and probes with their orientation information along with the Tm and locations. Then we group similar strings together and give them a group id. We used a simple logic, when we extract a string whose starting point is not between start and end points of previous key string (first string of a group), that string becomes a key string. There may be better ideas than this for forming groups. Because we used tight Tm ranges in this demo study, Figure 11 shows a small number of strings in every group, however number of oligonucleotides in a group can be high up to hundreds. Grouping is done separately for primers and probes, also separately for orientations.



```
*********
primer_a 312 TTTTACGGCAAGGCTATCCC / bas 4464 bit 4483 /  tm 59.38 / uctm -9.29 uc_100 46.86 / hmtm -19.61 hm_100 77.81
primer_a 312 TTTTACGGCAAGGCTATCCC / bas 4464 bit 4484 / tm 59.38318951404477
primer_a 312 TTTTACGGCAAGGCTATCCCC / bas 4464 bit 4485 / tm 61.861442348738535
primer_a 312 TTTACGGCAAGGCTATCCC / bas 4465 bit 4484 / tm 58.62917205547154
primer_a 312 TTTACGGCAAGGCTATCCCC / bas 4465 bit 4485 / tm 61.26361671046129
primer_a 312 TTACGGCAAGGCTATCCCC / bas 4466 bit 4485 / tm 60.60353833965297
primer_a 312 TACGGCAAGGCTATCCCCC / bas 4467 bit 4486 / tm 62.59879023356319
*********
probe_a 313 TTTTACGGCAAGGCTATCCCCCT / bas 4464 bit 4486 /  tm 65.40 / uctm -59.31 uc_100 63.17 / hmtm -19.61 hm_100 77.81
probe_a 313 TTTTACGGCAAGGCTATCCCCCT / bas 4464 bit 4487 / tm 65.40343040842356
probe_a 313 TTTACGGCAAGGCTATCCCCCT / bas 4465 bit 4487 / tm 65.010872917624
*********
primer_b 314 GGGGGATAGCCTTGCCGTA / bas 4467 bit 4485 /  tm 62.60 / uctm -151.18 uc_100 0.00 / hmtm 2.92 hm_100 69.03
primer_b 314 GGGGGATAGCCTTGCCGTA / bas 4466 bit 4485 / tm 62.59879023356319
primer_b 314 GGGGATAGCCTTGCCGTAA / bas 4465 bit 4484 / tm 60.60353833965297
*********
probe_b 315 AGGGGGATAGCCTTGCCGTAAA / bas 4465 bit 4486 /  tm 65.01 / uctm -34.98 uc_100 29.18 / hmtm 2.92 hm_100 69.03
*********
primer_b 316 CGAGCTCGTCGCACTTCTT / bas 4533 bit 4551 /  tm 61.81 / uctm -24.06 uc_100 40.05 / hmtm 35.90 hm_100 83.03
*********
primer_a 317 GTGATAGACTGCAACACGTG / bas 4692 bit 4711 /  tm 58.23 / uctm 24.61 uc_100 81.62 / hmtm 24.61 hm_100 81.62
primer_a 317 GTGATAGACTGCAACACGTG / bas 4692 bit 4712 / tm 58.23151549992093
primer_a 317 GTGATAGACTGCAACACGTGT / bas 4692 bit 4713 / tm 60.00177975884486
primer_a 317 GTGATAGACTGCAACACGTGTG / bas 4692 bit 4714 / tm 61.24115527106221
primer_a 317 GTGATAGACTGCAACACGTGTGT / bas 4692 bit 4715 / tm 62.77131194234579
primer_a 317 TGATAGACTGCAACACGTGT / bas 4693 bit 4713 / tm 58.60289186257228
primer_a 317 TGATAGACTGCAACACGTGTG / bas 4693 bit 4714 / tm 60.00188953183289
primer_a 317 TGATAGACTGCAACACGTGTGT / bas 4693 bit 4715 / tm 61.66010313894799
primer_a 317 TGATAGACTGCAACACGTGTGTC / bas 4693 bit 4716 / tm 62.52621527313107
primer_a 317 GATAGACTGCAACACGTGTG / bas 4694 bit 4714 / tm 58.23151549992093
primer_a 317 GATAGACTGCAACACGTGTGT / bas 4694 bit 4715 / tm 60.00177975884486
primer_a 317 GATAGACTGCAACACGTGTGTC / bas 4694 bit 4716 / tm 60.98844989618618
primer_a 317 GATAGACTGCAACACGTGTGTCA / bas 4694 bit 4717 / tm 62.52621527313107
primer_a 317 ATAGACTGCAACACGTGTGT / bas 4695 bit 4715 / tm 58.889564315597
primer_a 317 ATAGACTGCAACACGTGTGTC / bas 4695 bit 4716 / tm 60.00177975884486
primer_a 317 ATAGACTGCAACACGTGTGTCA / bas 4695 bit 4717 / tm 61.66010313894799
primer_a 317 ATAGACTGCAACACGTGTGTCAC / bas 4695 bit 4718 / tm 62.77131194234579
primer_a 317 TAGACTGCAACACGTGTGT / bas 4696 bit 4715 / tm 58.34995925705749
primer_a 317 TAGACTGCAACACGTGTGTC / bas 4696 bit 4716 / tm 59.54496453183447
primer_a 317 TAGACTGCAACACGTGTGTCA / bas 4696 bit 4717 / tm 61.289255246174605
primer_a 317 TAGACTGCAACACGTGTGTCAC / bas 4696 bit 4718 / tm 62.466407804110986
primer_a 317 AGACTGCAACACGTGTGTC / bas 4697 bit 4716 / tm 59.705884160370715
primer_a 317 AGACTGCAACACGTGTGTCA / bas 4697 bit 4717 / tm 61.52579853819623
primer_a 317 AGACTGCAACACGTGTGTCAC / bas 4697 bit 4718 / tm 62.740345519082325
primer_a 317 GACTGCAACACGTGTGTCA / bas 4698 bit 4717 / tm 60.00087110810102
primer_a 317 GACTGCAACACGTGTGTCAC / bas 4698 bit 4718 / tm 61.346123691995274
primer_a 317 ACTGCAACACGTGTGTCAC / bas 4699 bit 4718 / tm 60.2940245125577l
*********
```

Figure 13: Extracted oligonucleotides

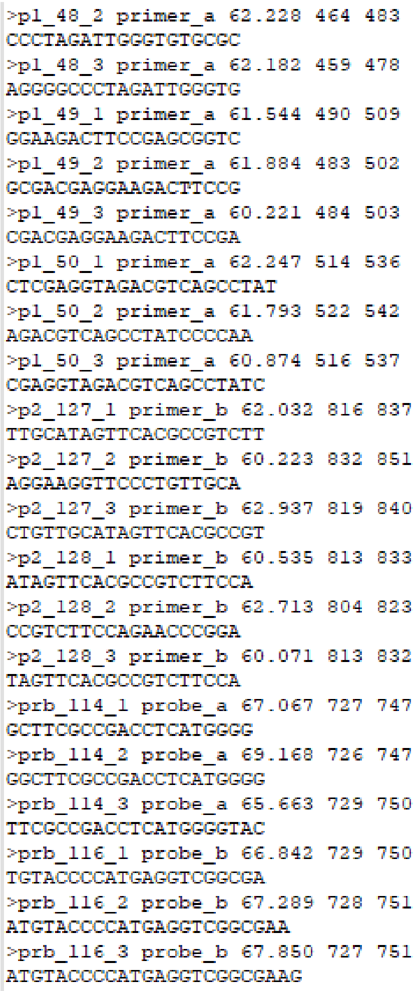
Then very carefully with start and end locations, we construct possible amplicons considering different constraints such as maximum allowed primer Tm difference, probe-primer Tm difference, amplicon length and some dimer constraints.



```
4731-4749---4878-4900---5008-5026  uzunluk 295 // probe1 // 321 339 347 // 58.97 65.75 58.68
4731-4749---4901-4921---5008-5026  uzunluk 295 // probe1 // 321 340 347 // 58.97 65.61 58.68
4731-4749---4888-4910---4914-4933  uzunluk 202 // probe2 // 321 344 341 // 58.97 65.84 59.32
4731-4749---4888-4910---5008-5026  uzunluk 295 // probe2 // 321 344 347 // 58.97 65.84 58.68
4731-4749---4911-4933---5008-5026  uzunluk 295 // probe2 // 321 343 347 // 58.97 65.31 58.68
4752-4770---4803-4820---4838-4856  uzunluk 104 // probe1 // 323 329 332 // 59.77 66.58 62.76
4752-4770---4803-4820---4914-4933  uzunluk 181 // probe1 // 323 329 341 // 59.77 66.58 59.32
4752-4770---4803-4820---4894-4912  uzunluk 160 // probe1 // 323 329 342 // 59.77 66.58 62.26
4752-4770---4803-4820---5008-5026  uzunluk 274 // probe1 // 323 329 347 // 59.77 66.58 58.68
4752-4770---4803-4820---5065-5083  uzunluk 331 // probe1 // 323 329 351 // 59.77 66.58 59.60
4752-4770---4821-4845---4914-4933  uzunluk 181 // probe1 // 323 330 341 // 59.77 65.15 59.32
4752-4770---4821-4845---4894-4912  uzunluk 160 // probe1 // 323 330 342 // 59.77 65.15 62.26
4752-4770---4821-4845---5008-5026  uzunluk 274 // probe1 // 323 330 347 // 59.77 65.15 58.68
4752-4770---4821-4845---5065-5083  uzunluk 331 // probe1 // 323 330 351 // 59.77 65.15 59.60
4752-4770---4818-4838---4838-4856  uzunluk 104 // probe2 // 323 336 332 // 59.77 65.19 62.76
4752-4770---4818-4838---4914-4933  uzunluk 181 // probe2 // 323 336 341 // 59.77 65.19 59.32
4752-4770---4818-4838---4894-4912  uzunluk 160 // probe2 // 323 336 342 // 59.77 65.19 62.26
4752-4770---4818-4838---5008-5026  uzunluk 274 // probe2 // 323 336 347 // 59.77 65.19 58.68
4752-4770---4818-4838---5065-5083  uzunluk 331 // probe2 // 323 336 351 // 59.77 65.19 59.60
4752-4770---4846-4862---4914-4933  uzunluk 181 // probe1 // 323 331 341 // 59.77 69.29 59.32
4752-4770---4846-4862---4894-4912  uzunluk 160 // probe1 // 323 331 342 // 59.77 69.29 62.26
4752-4770---4846-4862---5008-5026  uzunluk 274 // probe1 // 323 331 347 // 59.77 69.29 58.68
4752-4770---4846-4862---5065-5083  uzunluk 331 // probe1 // 323 331 351 // 59.77 69.29 59.60
4752-4770---4839-4858---4914-4933  uzunluk 181 // probe2 // 323 335 341 // 59.77 65.48 59.32
4752-4770---4839-4858---4894-4912  uzunluk 160 // probe2 // 323 335 342 // 59.77 65.48 62.26
4752-4770---4839-4858---5008-5026  uzunluk 274 // probe2 // 323 335 347 // 59.77 65.48 58.68
4752-4770---4839-4858---5065-5083  uzunluk 331 // probe2 // 323 335 351 // 59.77 65.48 59.60
4752-4770---4859-4876---4914-4933  uzunluk 181 // probe2 // 323 334 341 // 59.77 67.25 59.32
4752-4770---4859-4876---4894-4912  uzunluk 160 // probe2 // 323 334 342 // 59.77 67.25 62.26
4752-4770---4859-4876---5008-5026  uzunluk 274 // probe2 // 323 334 347 // 59.77 67.25 58.68
4752-4770---4859-4876---5065-5083  uzunluk 331 // probe2 // 323 334 351 // 59.77 67.25 59.60
4752-4770---4878-4900---4914-4933  uzunluk 181 // probe1 // 323 339 341 // 59.77 65.75 59.32
4752-4770---4878-4900---5008-5026  uzunluk 274 // probe1 // 323 339 347 // 59.77 65.75 58.68
4752-4770---4878-4900---5065-5083  uzunluk 331 // probe1 // 323 339 351 // 59.77 65.75 59.60
4752-4770---4901-4921---5008-5026  uzunluk 274 // probe1 // 323 340 347 // 59.77 65.61 58.68
4752-4770---4901-4921---5065-5083  uzunluk 331 // probe1 // 323 340 351 // 59.77 65.61 59.60
4752-4770---4888-4910---4914-4933  uzunluk 181 // probe2 // 323 344 341 // 59.77 65.84 59.32
4752-4770---4888-4910---5008-5026  uzunluk 274 // probe2 // 323 344 347 // 59.77 65.84 58.68
4752-4770---4888-4910---5065-5083  uzunluk 331 // probe2 // 323 344 351 // 59.77 65.84 59.60
4752-4770---4911-4933---5008-5026  uzunluk 274 // probe2 // 323 343 347 // 59.77 65.31 58.68
4752-4770---4911-4933---5065-5083  uzunluk 331 // probe2 // 323 343 351 // 59.77 65.31 59.60
4752-4770---4998-5022---5065-5083  uzunluk 331 // probe1 // 323 346 351 // 59.77 65.81 59.60
4752-4770---5001-5026---5065-5083  uzunluk 331 // probe2 // 323 348 351 // 59.77 65.40 59.60
4819-4837---4846-4862---4914-4933  uzunluk 114 // probe1 // 327 331 341 // 62.55 69.29 59.32
4819-4837---4846-4862---4894-4912  uzunluk 93 // probe1 // 327 331 342 // 62.55 69.29 62.26
4819-4837---4846-4862---5008-5026  uzunluk 207 // probe1 // 327 331 347 // 62.55 69.29 58.68
4819-4837---4846-4862---5065-5083  uzunluk 264 // probe1 // 327 331 351 // 62.55 69.29 59.60
4819-4837---4846-4862---5104-5122  uzunluk 303 // probe1 // 327 331 356 // 62.55 69.29 62.70
4819-4837---4839-4858---4914-4933  uzunluk 114 // probe2 // 327 335 341 // 62.55 65.48 59.32
```

Figure 14: Constructed amplicons on seed genome

In the next important step, we choose a fixed number of oligonucleotides from every group and we look for amplicons that are formed by these oligonucleotides in both target and non-target genomes. Here is one important detail, we do not pick amplicons from all possible amplicons and then query those oligonucleotides in other genomes; but we pick oligonucleotides from previously formed oligonucleotide groups. The reason is the following: as seen in Figure 11 some oligonucleotide groups may have many more oligonucleotides. As an extreme example, one primer_a (orientation a) group and one primer_b group in close proximity allowed by amplicon length both may have 1 single string and those two groups would generate 1 amplicon. However other two groups having each 100 strings will generate 10000 amplicons. So, sampling randomly from amplicons may generate a very non-uniform search space. We could have sampled a fixed number of amplicons from combinations of groups and it would be needlessly complicated and it already is close to what we did. However, there may be other ideas. Now we have all oligonucleotides to be queried on all other genomes, target or non-target.



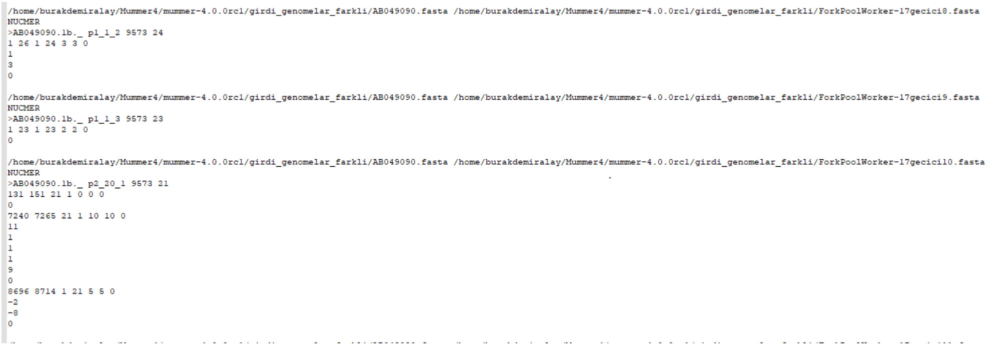
```
>p1_48_2 primer_a 62.228 464 483
CCCTAGATTGGGTGTGCGC
>p1_48_3 primer_a 62.182 459 478
AGGGGCCCTAGATTGGGTG
>p1_49_1 primer_a 61.544 490 509
GGAAGACTTCCGAGCGGTC
>p1_49_2 primer_a 61.884 483 502
GCGACGAGGAAGACTTCCG
>p1_49_3 primer_a 60.221 484 503
CGACGAGGAAGACTTCCGA
>p1_50_1 primer_a 62.247 514 536
CTCGAGGTAGACGTCAGCCTAT
>p1_50_2 primer_a 61.793 522 542
AGACGTCAGCCTATCCCCAA
>p1_50_3 primer_a 60.874 516 537
CGAGGTAGACGTCAGCCTATC
>p2_127_1 primer_b 62.032 816 837
TTGCATAGTTCACGCCGTCTT
>p2_127_2 primer_b 60.223 832 851
AGGAAGGTTCCCTGTTGCA
>p2_127_3 primer_b 62.937 819 840
CTGTTGCATAGTTCACGCCGT
>p2_128_1 primer_b 60.535 813 833
ATAGTTCACGCCGTCTTCCA
>p2_128_2 primer_b 62.713 804 823
CCGTCTTCCAGAACCCGGA
>p2_128_3 primer_b 60.071 813 832
TAGTTCACGCCGTCTTCCA
>prb_114_1 probe_a 67.067 727 747
GCTTCGCCGACCTCATGGGG
>prb_114_2 probe_a 69.168 726 747
GGCTTCGCCGACCTCATGGGG
>prb_114_3 probe_a 65.663 729 750
TTCGCCGACCTCATGGGGTAC
>prb_116_1 probe_b 66.842 729 750
TGTACCCCATGAGGTCGGCGA
>prb_116_2 probe_b 67.289 728 751
ATGTACCCCATGAGGTCGGCGAA
>prb_116_3 probe_b 67.850 727 751
ATGTACCCCATGAGGTCGGCGAAG
```

Figure 15: Chosen oligonucleotides to be queried

We now query every oligonucleotide separately to every single target and non-target genomes. This process is done parallel, every CPU studies one genome and switch to one other unprocessed genome when finished.

Now instead of the mummer command, we query with nucmer command that allows mismatches. We use the command: # nucmer —maxmatch -l minimum_length_of_anchor -c minimum_length_of_alignment –nooptimize –save location+id ref_file qry_file #. All variables are explained in detail in experimental results section. This part is the heart of the algorithm and the most important reason why our method is so successful.

Here we first generate dummy out files for every query string and then concatenate to one file for each genome.



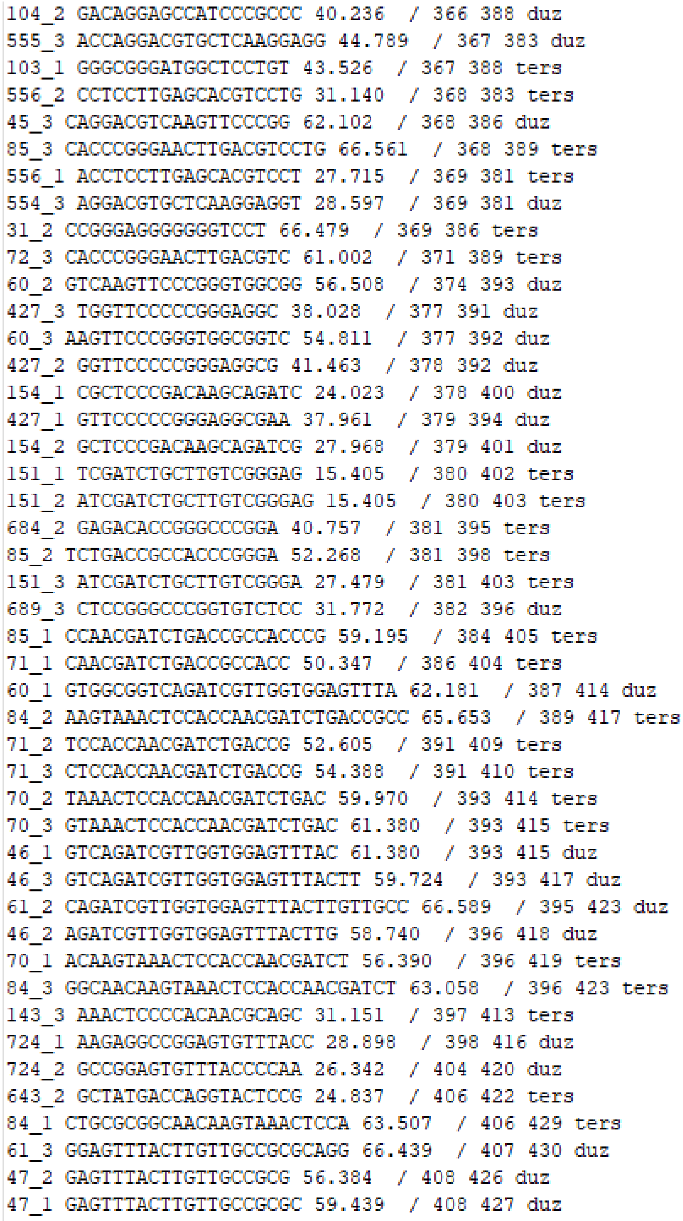
```
/home/burakdemiralay/Mummer4/mummer-4.0.0rc1/girdi_genomelar_farkli/AB049090.fasta /home/burakdemiralay/Mummer4/mummer-4.0.0rc1/girdi_genomelar_farkli/ForkPoolWorker-17gecici8.fasta
NUCMER
>AB049090.1b._ p1_1_2 9573 24
1 26 1 24 3 3 0
1
3
0

/home/burakdemiralay/Mummer4/mummer-4.0.0rc1/girdi_genomelar_farkli/AB049090.fasta /home/burakdemiralay/Mummer4/mummer-4.0.0rc1/girdi_genomelar_farkli/ForkPoolWorker-17gecici9.fasta
NUCMER
>AB049090.1b._ p1_1_3 9573 23
1 23 1 23 2 2 0
0

/home/burakdemiralay/Mummer4/mummer-4.0.0rc1/girdi_genomelar_farkli/AB049090.fasta /home/burakdemiralay/Mummer4/mummer-4.0.0rc1/girdi_genomelar_farkli/ForkPoolWorker-17gecici10.fasta
NUCMER
>AB049090.1b._ p2_20_1 9573 21
131 151 21 1 0 0 0
0
7240 7265 21 1 10 10 0
11
1
1
1
9
0
8696 8714 1 21 5 5 0
-2
-8
0
```

Figure 16: Suffix array result of queries

Explanation of output file shown in Figure 11 is as follows; 7240 7265 21 1 ~~10 10 0~~ # start location of hit in reference; end location of hit in reference; start location of hit in query, end location of hit in query. If end location of hit in query is smaller than start, then alignment is in complementary strand. Until we see a single 0 in a row, other numbers in rows mean; if >0; insertion in reference/deletion in query and <0; vice versa.

So, we now can extract hit locations, sometimes multiple hit locations for every oligonucleotide for every genome.

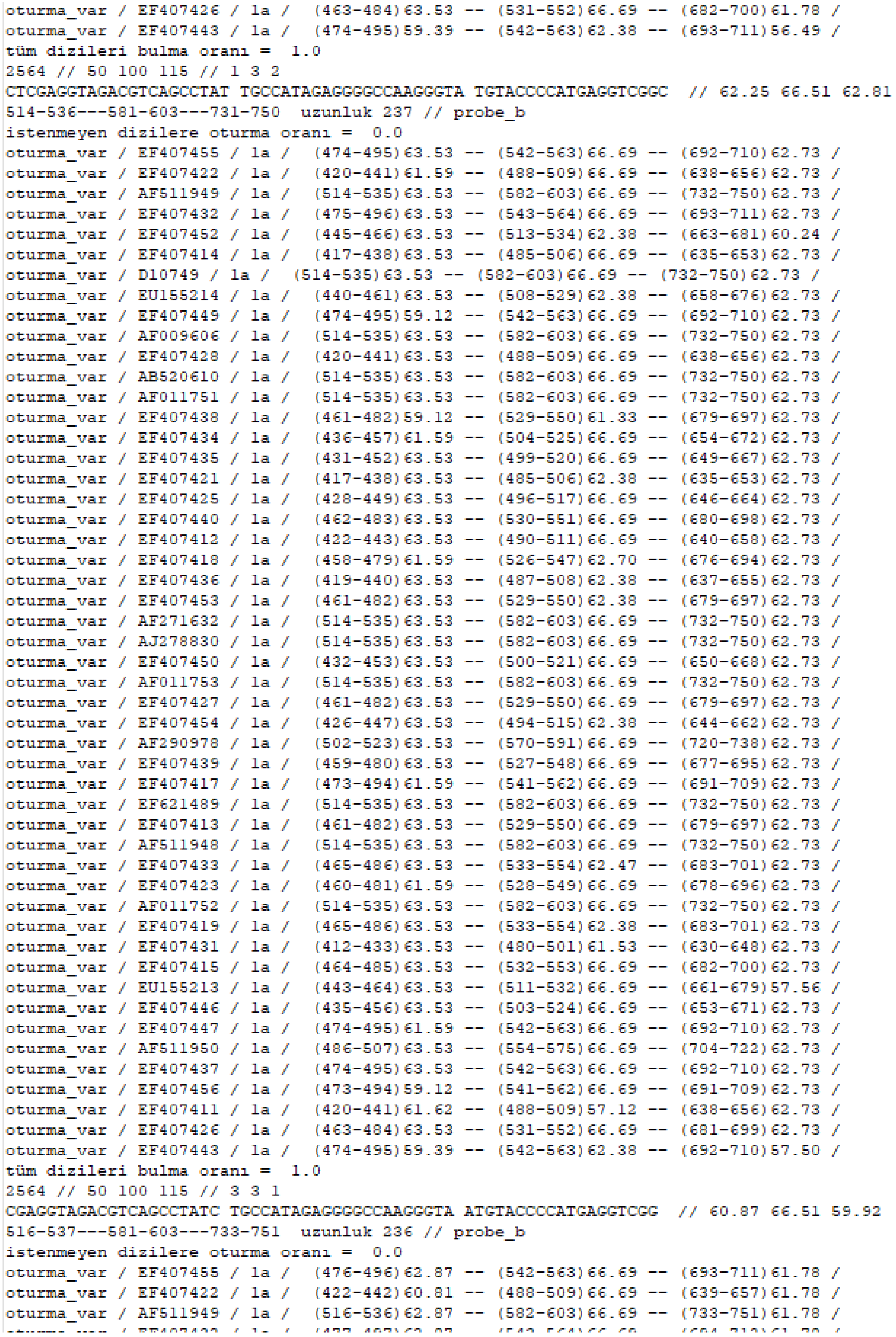
After extracting all hit locations for every oligonucleotide, we can look at their interaction thermodynamically, most important factor. We generate a file for every genome, target or non-target, that which oligonucleotide interacts at 0.5 probability at what temperature and locations and to which strand. This process is also fully parallel and every CPU holds one genome and switches to next after finishing.



```
104_2 GACAGGAGCCATCCCGCCC 40.236  / 366 388 duz
555_3 ACCAGGACGTGCTCAAGGAGG 44.789  / 367 383 duz
103_1 GGGCGGGATGGCTCCTGT 43.526  / 367 388 ters
556_2 CCTCCTTGAGCACGTCCTG 31.140  / 368 383 ters
45_3 CAGGACGTCAAGTTCCCGG 62.102  / 368 386 duz
85_3 CACCCGGGAACTTGACGTCCTG 66.561  / 368 389 ters
556_1 ACCTCCTTGAGCACGTCCT 27.715  / 369 381 ters
554_3 AGGACGTGCTCAAGGAGGT 28.597  / 369 381 duz
31_2 CCGGGAGGGGGGGTCCT 66.479  / 369 386 ters
72_3 CACCCGGGAACTTGACGTC 61.002  / 371 389 ters
60_2 GTCAAGTTCCCGGGTGGCGG 56.508  / 374 393 duz
427_3 TGGTTCCCCCGGGAGGC 38.028  / 377 391 duz
60_3 AAGTTCCCGGGTGGCGGTC 54.811  / 377 392 duz
427_2 GGTTCCCCCGGGAGGCG 41.463  / 378 392 duz
154_1 CGCTCCCGACAAGCAGATC 24.023  / 378 400 duz
427_1 GTTCCCCCGGGAGGCGAA 37.961  / 379 394 duz
154_2 GCTCCCGACAAGCAGATCG 27.968  / 379 401 duz
151_1 TCGATCTGCTTGTCGGGAG 15.405  / 380 402 ters
151_2 ATCGATCTGCTTGTCGGGAG 15.405  / 380 403 ters
684_2 GAGACACCGGGCCCGGA 40.757  / 381 395 ters
85_2 TCTGACCGCCACCCGGGA 52.268  / 381 398 ters
151_3 ATCGATCTGCTTGTCGGGA 27.479  / 381 403 ters
689_3 CTCCGGGCCCGGTGTCTCC 31.772  / 382 396 duz
85_1 CCAACGATCTGACCGCCACCCG 59.195  / 384 405 ters
71_1 CAACGATCTGACCGCCACC 50.347  / 386 404 ters
60_1 GTGGCGGTCAGATCGTTGGTGGAGTTTA 62.181  / 387 414 duz
84_2 AAGTAAACTCCACCAACGATCTGACCGCC 65.653  / 389 417 ters
71_2 TCCACCAACGATCTGACCG 52.605  / 391 409 ters
71_3 CTCCACCAACGATCTGACCG 54.388  / 391 410 ters
70_2 TAAACTCCACCAACGATCTGAC 59.970  / 393 414 ters
70_3 GTAAACTCCACCAACGATCTGAC 61.380  / 393 415 ters
46_1 GTCAGATCGTTGGTGGAGTTTAC 61.380  / 393 415 duz
46_3 GTCAGATCGTTGGTGGAGTTTACTT 59.724  / 393 417 duz
61_2 CAGATCGTTGGTGGAGTTTACTTGTTGCC 66.589  / 395 423 duz
46_2 AGATCGTTGGTGGAGTTTACTTG 58.740  / 396 418 duz
70_1 ACAAGTAAACTCCACCAACGATCT 56.390  / 396 419 ters
84_3 GGCAACAAGTAAACTCCACCAACGATCT 63.058  / 396 423 ters
143_3 AAACTCCCCACAACGCAGC 31.151  / 397 413 ters
724_1 AAGAGGCCGGAGTGTTTACC 28.898  / 398 416 duz
724_2 GCCGGAGTGTTTACCCCAA 26.342  / 404 420 duz
643_2 GCTATGACCAGGTACTCCG 24.837  / 406 422 ters
84_1 CTGCGCGGCAACAAGTAAACTCCA 63.507  / 406 429 ters
61_3 GGAGTTTACTTGTTGCCGCGCAGG 66.439  / 407 430 duz
47_2 GAGTTTACTTGTTGCCGCG 56.384  / 408 426 duz
47_1 GAGTTTACTTGTTGCCGCGC 59.439  / 408 427 duz
```

Figure 17: Thermodynamic interaction results of oligonucleotides with an input genome

Now in the final step, according to desired true and false positive rates, we output a final report file that contains every single amplicon in every genome if there is an amplification *in silico*. This process is also parallelized on every CPU and there are even multiple threads on every CPU because this process is very reading and writing intensive.



```
oturma_var / EF407426 / la /  (463-484)63.53 -- (531-552)66.69 -- (682-700)61.78 /
oturma_var / EF407443 / la /  (474-495)59.39 -- (542-563)62.38 -- (693-711)56.49 /
tüm dizileri bulma oranı =  1.0
2564 // 50 100 115 // 1 3 2
CTCGAGGTAGACGTCAGCCTAT TGCCATAGAGGGGCCAAGGGTA TGTACCCCATGAGGTCGGC  // 62.25 66.51 62.81
514-536---581-603---731-750  uzunluk 237 // probe_b
istenmeyen dizilere oturma oranı =  0.0
oturma_var / EF407455 / la /  (474-495)63.53 -- (542-563)66.69 -- (692-710)62.73 /
oturma_var / EF407422 / la /  (420-441)61.59 -- (488-509)66.69 -- (638-656)62.73 /
oturma_var / AF511949 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407432 / la /  (475-496)63.53 -- (543-564)66.69 -- (693-711)62.73 /
oturma_var / EF407452 / la /  (445-466)63.53 -- (513-534)62.38 -- (663-681)60.24 /
oturma_var / EF407414 / la /  (417-438)63.53 -- (485-506)66.69 -- (635-653)62.73 /
oturma_var / D10749 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EU155214 / la /  (440-461)63.53 -- (508-529)62.38 -- (658-676)62.73 /
oturma_var / EF407449 / la /  (474-495)59.12 -- (542-563)66.69 -- (692-710)62.73 /
oturma_var / AF009606 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407428 / la /  (420-441)63.53 -- (488-509)66.69 -- (638-656)62.73 /
oturma_var / AB520610 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / AF011751 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407438 / la /  (461-482)59.12 -- (529-550)61.33 -- (679-697)62.73 /
oturma_var / EF407434 / la /  (436-457)61.59 -- (504-525)66.69 -- (654-672)62.73 /
oturma_var / EF407435 / la /  (431-452)63.53 -- (499-520)66.69 -- (649-667)62.73 /
oturma_var / EF407421 / la /  (417-438)63.53 -- (485-506)62.38 -- (635-653)62.73 /
oturma_var / EF407425 / la /  (428-449)63.53 -- (496-517)66.69 -- (646-664)62.73 /
oturma_var / EF407440 / la /  (462-483)63.53 -- (530-551)66.69 -- (680-698)62.73 /
oturma_var / EF407412 / la /  (422-443)63.53 -- (490-511)66.69 -- (640-658)62.73 /
oturma_var / EF407418 / la /  (458-479)61.59 -- (526-547)62.70 -- (676-694)62.73 /
oturma_var / EF407436 / la /  (419-440)63.53 -- (487-508)62.38 -- (637-655)62.73 /
oturma_var / EF407453 / la /  (461-482)63.53 -- (529-550)62.38 -- (679-697)62.73 /
oturma_var / AF271632 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / AJ278830 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407450 / la /  (432-453)63.53 -- (500-521)66.69 -- (650-668)62.73 /
oturma_var / AF011753 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407427 / la /  (461-482)63.53 -- (529-550)66.69 -- (679-697)62.73 /
oturma_var / EF407454 / la /  (426-447)63.53 -- (494-515)62.38 -- (644-662)62.73 /
oturma_var / AF290978 / la /  (502-523)63.53 -- (570-591)66.69 -- (720-738)62.73 /
oturma_var / EF407439 / la /  (459-480)63.53 -- (527-548)66.69 -- (677-695)62.73 /
oturma_var / EF407417 / la /  (473-494)61.59 -- (541-562)66.69 -- (691-709)62.73 /
oturma_var / EF621489 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407413 / la /  (461-482)63.53 -- (529-550)66.69 -- (679-697)62.73 /
oturma_var / AF511948 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407433 / la /  (465-486)63.53 -- (533-554)62.47 -- (683-701)62.73 /
oturma_var / EF407423 / la /  (460-481)61.59 -- (528-549)66.69 -- (678-696)62.73 /
oturma_var / AF011752 / la /  (514-535)63.53 -- (582-603)66.69 -- (732-750)62.73 /
oturma_var / EF407419 / la /  (465-486)63.53 -- (533-554)62.38 -- (683-701)62.73 /
oturma_var / EF407431 / la /  (412-433)63.53 -- (480-501)61.53 -- (630-648)62.73 /
oturma_var / EF407415 / la /  (464-485)63.53 -- (532-553)66.69 -- (682-700)62.73 /
oturma_var / EU155213 / la /  (443-464)63.53 -- (511-532)66.69 -- (661-679)57.56 /
oturma_var / EF407446 / la /  (435-456)63.53 -- (503-524)66.69 -- (653-671)62.73 /
oturma_var / EF407447 / la /  (474-495)61.59 -- (542-563)66.69 -- (692-710)62.73 /
oturma_var / AF511950 / la /  (486-507)63.53 -- (554-575)66.69 -- (704-722)62.73 /
oturma_var / EF407437 / la /  (474-495)63.53 -- (542-563)66.69 -- (692-710)62.73 /
oturma_var / EF407456 / la /  (473-494)59.12 -- (541-562)66.69 -- (691-709)62.73 /
oturma_var / EF407411 / la /  (420-441)61.62 -- (488-509)57.12 -- (638-656)62.73 /
oturma_var / EF407426 / la /  (463-484)63.53 -- (531-552)66.69 -- (681-699)62.73 /
oturma_var / EF407443 / la /  (474-495)59.39 -- (542-563)62.38 -- (692-710)57.50 /
tüm dizileri bulma oranı =  1.0
2564 // 50 100 115 // 3 3 1
CGAGGTAGACGTCAGCCTATC TGCCATAGAGGGGCCAAGGGTA ATGTACCCCATGAGGTCGG  // 60.87 66.51 59.92
516-537---581-603---733-751  uzunluk 236 // probe_b
istenmeyen dizilere oturma oranı =  0.0
oturma_var / EF407455 / la /  (476-496)62.87 -- (542-563)66.69 -- (693-711)61.78 /
oturma_var / EF407422 / la /  (422-442)60.81 -- (488-509)66.69 -- (639-657)61.78 /
oturma_var / AF511949 / la /  (516-536)62.87 -- (582-603)66.69 -- (733-751)61.78 /
```

Figure 18: End Result

APPENDIX C

HPV In Vitro Study Results

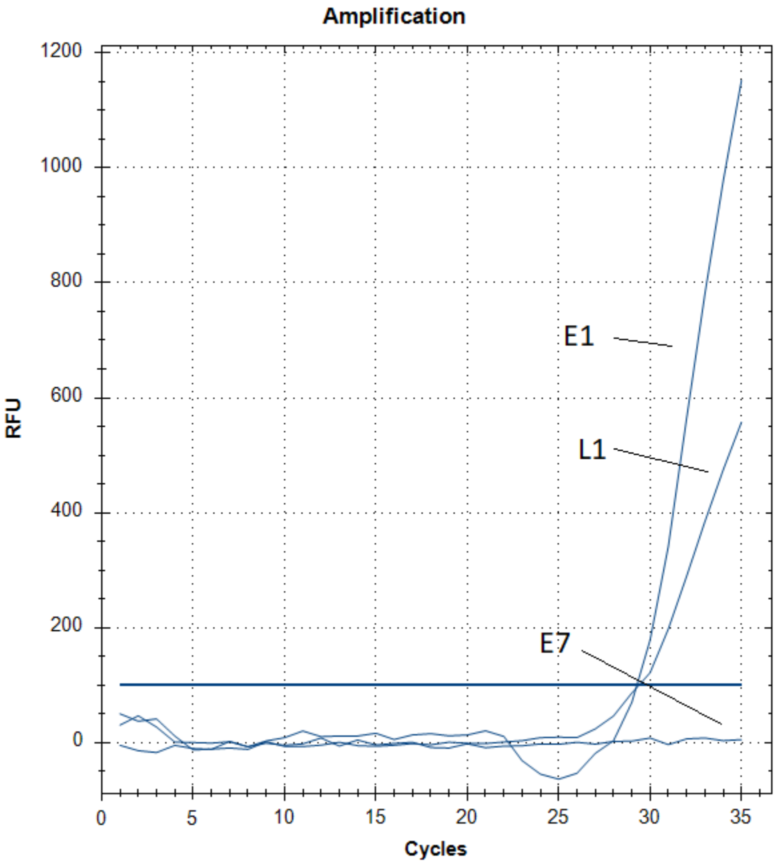


Figure 19: Positive Control Sample, different oligonucleotide sets of HPV-16

Figure 19 shows overlay of three different PCR wells in same run. Because HPV-16 is the highest risk HPV subtype, we wanted to be sure of its amplification and signal quality, so we designed three oligonucleotides set targeting three different regions of HPV. Sample is positive control mixture of HPV-16, HPV-18 and HPV-45.

Here there is no amplification from E7 region and we see better amplification from E1 compared to L1. This shows importance of virus dynamics information before designing primers and probes.

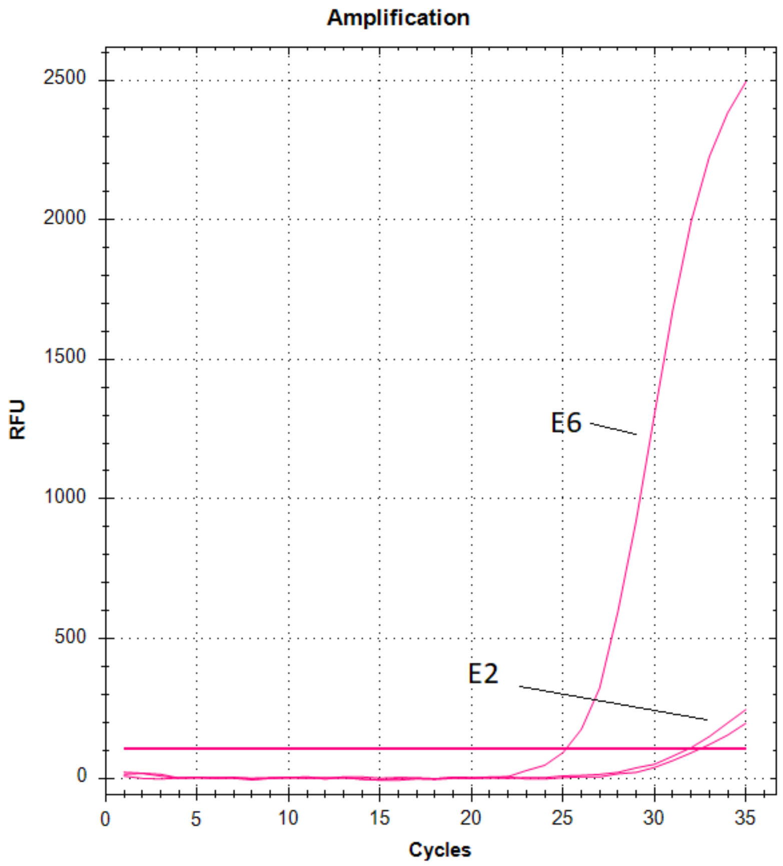


Figure 20: Positive Control Sample, different oligonucleotide sets of HPV-45

Figure 20 shows overlay of three different PCR wells in same run. Sample is positive control mixture of HPV-16, HPV-18 and HPV-45.

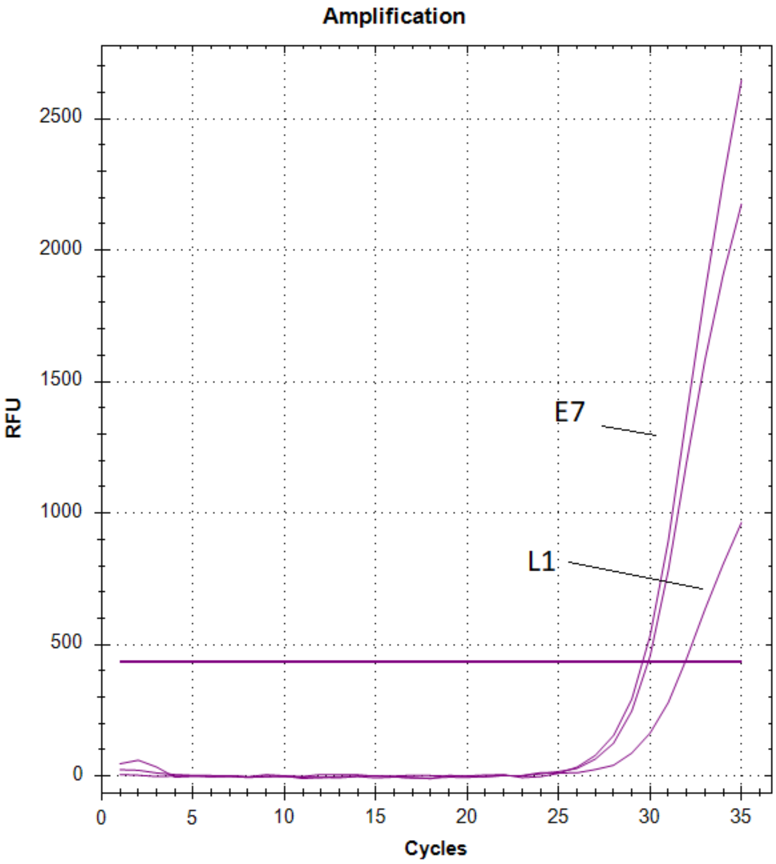


Figure 21: Positive Control Sample, different oligonucleotide sets of HPV-18

Figure 21 shows overlay of three different PCR wells in same run. Sample is positive control mixture of HPV-16, HPV-18 and HPV-45.

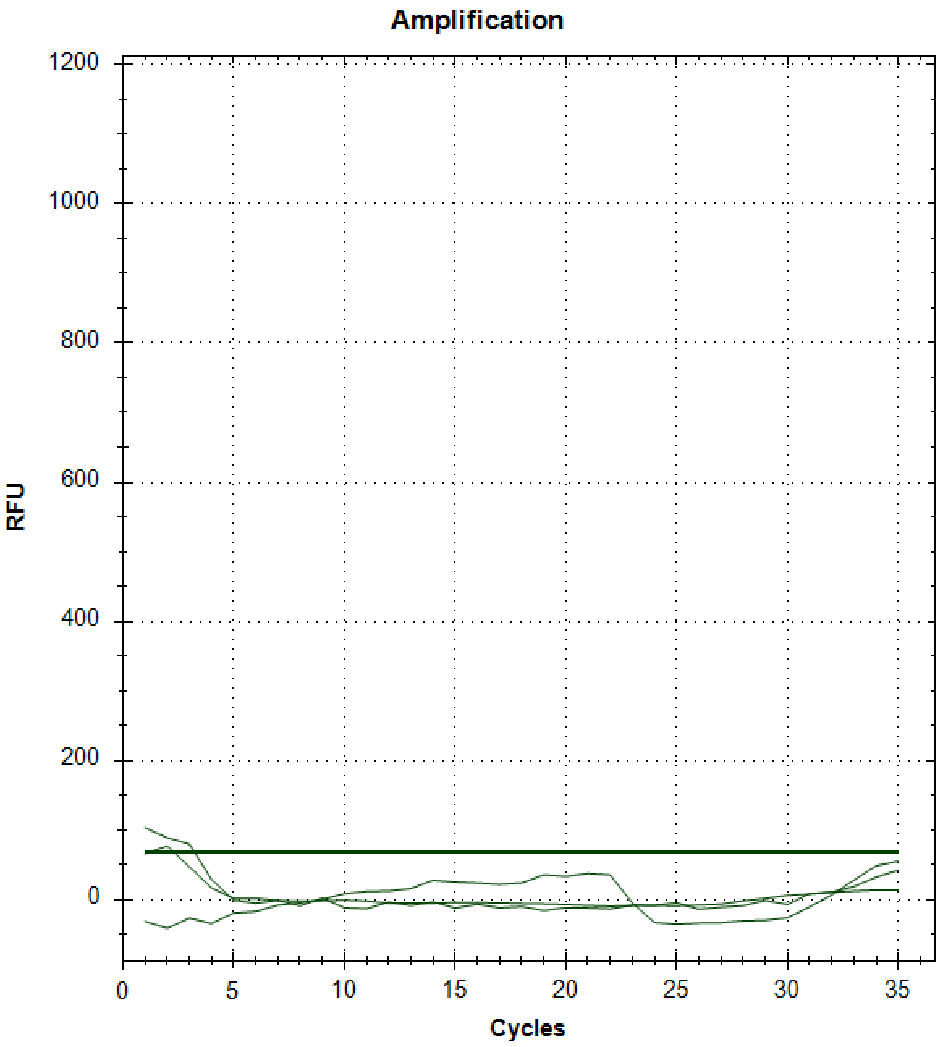


Figure 22: Positive Control Sample, oligonucleotide set of HPV-39-59-68

Figure 22 shows overlay of three different PCR wells in same run. Sample is positive control mixture of HPV-16, HPV-18 and HPV-45. We used oligonucleotide set targeting E1 region that is uniquely conserved in all of HPV 39, HPV-59 and HPV 68. Our expectation was not to observe any amplification.

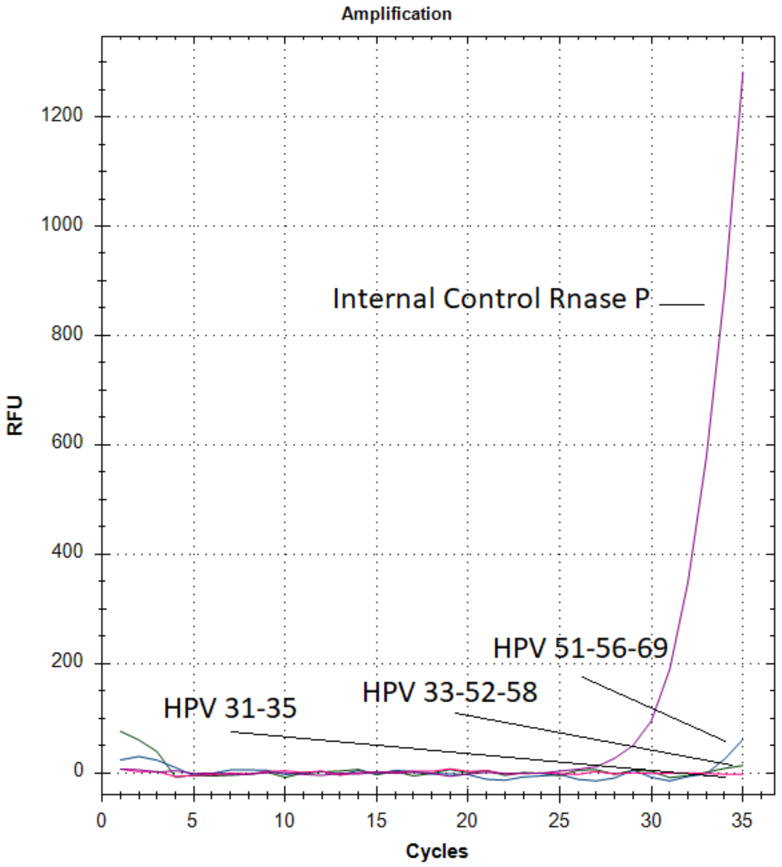


Figure 23: Positive Control Sample, corresponding oligonucleotide sets of Tube 2

Figure 23 shows positive control mixture of HPV-16, HPV-18 and HPV-45. We used oligonucleotide sets differentially targeting E1 regions of HPV-51-56-69, HPV 33-52-58 and HPV 31-35. Only internal control is amplified as expected.

APPENDIX D

Supplementary Data

We report all results in high resolution in Supplementary Data. For common region and differentiation studies of HIV, HCV and Dengue Virus; files named as Supplementary_virusname_subtype.txt files include NCBI id of genomes sequenced, subtypes they belong to, location of oligonucleotides in that specific genome and hybridization melting temperatures with that region and whether it is an accepted interaction or not. Files also include chosen oligonucleotides, their locations in seed genome, false positive and true positive rates. In case of false positives, every genome individually is reported with mentioned properties.

We also report the results of the sequence parameter analysis on Dengue Virus and HIV. These files are named as Supplementary_virusname_lengthofseedgenome_lengthofalignment.txt. They include every found oligonucleotide set with the mentioned properties for common regions and differentiation studies based on the chosen seed genome length and alignment length. We have included result files comparing our method with PrimerHunter, which we tested for distinguishing between HCV subtype 1a and subtype 1b.

All output result files and all raw input genome files, are provided in Supplementary Data and at the following link: https://github.com/Burak1Demiralay

CURRICULUM VITAE

PERSONAL INFORMATION

Demiralay Burak

EDUCATION

PhD	Health Informatics, METU	2024
MSc	Bionformatics, METU	2012
BSc	Molecular Biology and Genetics, METU	2009
High School	Ankara Fen lisesi	1999

INTERESTS

Satisfiability Problem

Bioinformatics