P.PORTO

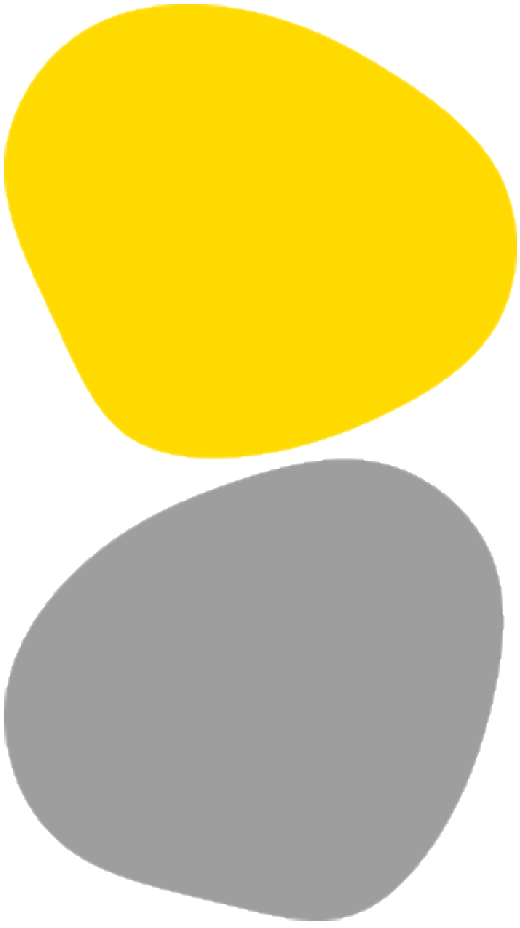
ESCOLA SUPERIOR DE SAÚDE

M MESTRADO

BIOESTATÍSTICA E BIOINFORMÁTICA APLICADAS À SAÚDE

ViruScopeDB: a comprehensive multi-omics software and database for highly infectious viruses

Ana Sofia Fafiães Pires de Lima



09/**2025**

ViruScopeDB: a comprehensive multi-omics software and database for highly infectious viruses

Author

Ana Sofia Fafiães Pires de Lima

Supervisors

PhD / João Carneiro / Interdisciplinary Center of Marine and Environmental Research (CIIMAR)

PhD / Sérgio Sousa / LAQV/REQUIMTE - BioSIM, Department of Biomedicine, Faculty of Medicine, University of Porto (FMUP)

PhD / Vítor Sá / LabRP/CIR, School of Health, Polytechnic of Porto (E2S IPP)

*Dissertation presented to fulfill the requirements necessary to obtain the Master's degree in **Biostatistics and Bioinformatics Applied to Health** by the School of Health of the Polytechnic Institute of Porto.*

Funding

This research was supported by Portuguese national funds through the Foundation for Science and Technology (FCT) within the scope of UIDB/04423/2020 (CIIMAR) and UIDP/04423/2020 (CIIMAR), 10.54499/LA/P/0008/2020,10.54499/UIDP/50006/2020, UIDB/04050/2020 (UMinho CBMA - https://doi.org/10.54499/UIDB/04050/2020), 10.54499/UIDB/50006/2020 (LAVQ), UIDB/00319/2020 (ALGORITMI/LASI) and UIDB/00127/2020 (IEETA/LASI - doi.org/10.54499/UIDB/00127/2020). JC acknowledges the FCT funding for his research contract established under the transitional rule of Decree Law 57/2016, amended by Law 57/2017.

The Galaxy server used for some calculations is partly funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.

fct Fundação para a Ciência e a Tecnologia

REPÚBLICA PORTUGUESA

Acknowledgements

I would like to begin by expressing my deepest gratitude to my supervisors, who have shown me incredible patience, guidance, and unwavering support throughout the duration of this project. My growth as a researcher, bioinformatician and all-around professional this past year is a consequence of the invaluable knowledge they have imparted upon me, and the doors they have opened along the way.

To Dr. João Carneiro, I thank you for giving me the opportunity to work on a project that connected with me, for always encouraging me to chase my ideas, and for kindly guiding me through the obstacles I faced along the way. The trust shown to me during this project greatly contributed to my growing confidence in myself as a member of the scientific community.

To Dr. Sérgio Sousa, I cannot overstate how insightful the discussions had over the past months have been. Your collaboration was key for the shaping and enrichment of the project.

To Dr. Vítor J. Sá, I am sincerely grateful for the enthusiasm and continuous encouragement given to me. The opportunities you have blessed me with, to share my hard work and passion for this area with the community and be met with positive reactions, were the boosts I needed to keep going when I felt like I was running out of steam.

Finally, to the unofficial supervisor of the project, Diogo Pratas, I would like to underscore my appreciation for the help given in bridging my Biology and Informatics skills, and the ever-present availability to help. I aspire to one day be a professional with the same balance of knowledge, experience and approachability as the one I got to witness this year.

This work also benefitted from the support of the teams at CIIMAR, BioSIM, E2S, as well as fellow students conducting their research at these institutions in tandem with me. The warm welcomes and mutual support were vital for the success of my endeavors.

On a more personal note, I am forever grateful to my family, particularly and most importantly my parents, who have shared the ups and downs of my academic career all these years. The person I am today is a result of the unconditional nurture, love, and care I have always received from you. To my boyfriend, Andrey, who cheered me on every day, who listened to me complain about the same things over and over,

and who saved my entire project that one time, I am so lucky to have you by my side, and I can only hope to be half the support you were to me if or when the roles reverse.

Once again, to my parents, partner ~~and cat~~, who supported me through my temporary defeats, my periods of demotivation and feeling overwhelmed: I am so proud to share this victory with you. It's yours as much as it is mine.

Resumo

Vírus altamente infeciosos, como o HIV, Ebolavírus e SARS-CoV-2, têm apresentado desafios contínuos para a saúde global, deixando no seu rasto consequências devastadoras. Estas destabilizações não se limitaram apenas ao ramo da saúde, provocando sequelas a nível económico e social que impactaram, e continuarão a ter impacto na sociedade durante décadas. Monitorização eficaz e testes de diagnóstico rápidos continuam a ser fatores essenciais para o controlo da propagação, no entanto os atuais métodos para coleção e curadoria de dados relativos a oligonucleótidos são fragmentados, direcionados para vírus específicos, e geralmente dependentes de trabalho manual. De forma a colmatar estas limitações, esta tese detalha a criação do ViruScope, uma ferramenta bioinformática automatizada para *data mining* de informação relativa às -ómicas, geração, validação e avaliação de primers *in silico*, junto com a ViruScopeDB, uma base de dados centralizada, atualizada e multi-viral, construída com os resultados gerados pela ferramenta.

A metodologia integra a recolha e alinhamento de dados genómicos a grande escala, extração de primers experimentalmente validados da literatura (AROLit v.3), geração e filtragem de primers gerados *in silico* (iSOP v.3), e uma nova abordagem para classificação de primers, que engloba parâmetros como a percentagem de conteúdo GC, temperatura de fusão, conservação inter-espécies, e um novo algoritmo para ordenar primers em relação à probabilidade do mesmo de formar estruturas secundárias indesejáveis (No-Fold %). Estas funções, desenhadas anteriormente para vírus específicos, foram reescritas para formatos escaláveis, otimizados e generalizados, e lançados em formato de programa para a linha de comando, e em forma de aplicação com interface gráfica com o Shiny, enquanto que a base de dados foi criada como um repositório aberto.

Casos de estudo utilizando dados relativos ao Ébola demonstram a capacidade da ferramenta de recolher, validar e classificar primers tanto da literatura como gerados computacionalmente, produzindo resultados consistentes com práticas experimentais, com a vantagem acrescida de poder ser feito o sorteio relativo das sequências. Filtrar o conjunto de dados pelos valores de conservação e No-Fold% - parâmetros finais e exclusivos do ViruScope – revelaram diferenças a nível da conservação e estabilidade relacionadas com a posição no genoma, reforçando a vantagem de métodos automatizados e escaláveis para o design de ferramentas de diagnóstico.

Ao uniformizar a recolha de dados, design de primers e validação para uma *pipeline*, o ViruScope e a ViruScopeDB fornecem, assim, uma nova solução para alguns dos desafios de diagnóstico molecular, seguindo sempre a ideologia FAIR (Findable, Accessible, Interoperable, Reusable). Desta forma, o presente trabalho demonstra o potencial de *pipelines* computacionais para a aceleração de novas estratégias de diagnóstico, melhoria da monitorização de vírus diversos, e para reforçar a preparação global tanto para ameaças atuais, como para potenciais futuras crises.

Palavras-chave: bioinformática, virologia, base de dados, PCR, oligonucleótidos.

Abstract

Highly infectious viruses, such as HIV, Ebolavirus, and SARS-CoV-2 have presented ongoing challenges to global health, leaving devastating consequences in their wake. These disruptions were not limited to the health field, as their profound economic and societal impact has been, and will continue to be, felt for decades. Effective surveillance and rapid diagnostics remain crucial to controlling their spread, yet current approaches to oligonucleotide data curation are fragmented, virus-specific, and generally highly dependent on manual approaches. To address these limitations, this thesis presents ViruScope, an automated bioinformatics tool for multi-omics data mining, *in silico* primer generation, validation, and scoring, together with ViruScopeDB, a centralized and continuously updated cross-viral database built on the results of the tool.

The methodology integrates large-scale genomic data retrieval and alignment, literature-based extraction of experimentally validated primers (AROLit v.3), *in silico* primer generation and filtering (iSOP v.3), and a novel scoring framework, encompassing GC content, melting temperature, cross-species conservation, and a way to rank primers by how likely they are to not form undesirable secondary structures (No-Fold%). These modules, previously designed for specific viruses, were re-engineered into a scalable, optimized and generalized pipeline, deployed as both a command-line program and a user-friendly Shiny application, and compiled into ViruScopeDB as an openly accessible repository.

Case studies conducted on Ebola datasets demonstrated the tool's capacity to retrieve, validate, and score primers from both literature and computational approaches, producing results consistent with experimental best practices while enabling more nuanced ranking of primer candidates. Filtering by conservation scores and No-Fold% – the final and ViruScope-exclusive parameters – revealed gene and region-specific differences in conservation and stability, reinforcing the advantage in automated and scalable approaches for diagnostic tool design.

By unifying data retrieval, primer design, and validation into a reproducible, automated workflow, ViruScope and ViruScopeDB provide a novel, scalable, and FAIR (Findable, Accessible, Interoperable, Reusable) solution to some of the challenges of molecular diagnostics. This work demonstrates the potential of computational pipelines to accelerate diagnostic assay development, improve cross-viral surveillance, and strengthen global preparedness for both ongoing and potential emerging viral threats.

Keywords: bioinformatics, virology, database, PCR, oligonucleotides.

Table of Contents

List of figures

List of tables

List of Equations

List of acronyms and abbreviations

AIDS – Acquired Immunodeficiency Syndrome

API – Application Programming Interface

APOBEC3 – Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3

AROLit – Automatic Retrieval of Oligonucleotides from Literature

BLAST – Basic Local Alignment Search Tool

Bp – base pairs

CDC – Centers for Disease Control and Prevention

cDNA – Complementary DNA

CDS – Coding DNA Sequence

CPU – Central Processing Unit

CRF – Circulating Recombinant Forms

CS – Conservation Score

DOI – Digital Object Identifier

E – Envelope

EHF – Ebola Hemorrhagic Fever

EVD – Ebola Virus Disease

FAIR – Findable, Accessible, Interoperable, Reusable

GDP – Gross Domestic Product

GP – Glycoprotein

GUI – Graphical User Interface

HIV – Human Immunodeficiency Virus

iSOP – *In Silico* Oligonucleotides designed in Python

IUPAC – International Union of Pure and Applied Chemistry

L – Ebolavirus RNA-dependent RNA Polymerase

LLM – Large Language Model

M – Membrane

MAFFT – Multiple Alignment using Fast Fourier Transform

MSA – Multiple Sequence Alignment

N – Nucleocapsid

NCBI – National Centre for Biotechnology Information

NP – Nucleoprotein

nsp – Non-structural Protein

OCR – Optical Character Recognition

PCR – Polymerase Chain Reaction

PIS – Proportion of Identical Sites

PPI – Percentage of Pairwise Identity

PPI3' – Percentage of Pairwise Identity of the last 3 nucleotides on the 3' end

RefSeq – Reference Sequence

RegEx – Regular Expressions

RT – Reverse Transcription

RT-PCR – Real-Time Polymerase Chain Reaction

S – Spike

SARS-CoV-2 – Severe Acute Respiratory Syndrome coronavirus 2

SNP – Single Nucleotide Polymorphism

ssRNA – Single-stranded RNA

SVG – Scalable Vector Graphics

Tm – Melting Temperature

VP – Viral Protein

WHO – World Health Organization

ΔG – Gibbs free energy difference

1. Introduction

Viral pathogens have long co-existed with humans, with some, like the diverse plethora of genera responsible for the common cold, even becoming endemic – meaning that its presence is prevalent over time in a population contained within a geographic area (*Epidemic, Endemic, Pandemic*, 2021). Others, like the Variola virus (commonly known as "smallpox"), faced a totally opposite fate – total eradication, defined as the reduction of the incidence of a disease to zero, on a global scale (Dowdle, 1998). However, the past decades have repeatedly proven the devastating consequences of highly infectious viruses on global health, especially when this heightened infectivity is accompanied by a similar degree of virulence. From the HIV/AIDS pandemic of the 1980s, still ongoing to this day, to the West Africa Ebola outbreak in 2014-2016, and, most recently, the 2020 COVID-19 pandemic, these pathogens have highlighted the fragility and unpreparedness of healthcare systems on a global scale, reflecting the urgent need for robust and scalable detection methods, as well as continuous surveillance programs, in order to mitigate the burdens of any current or future occurrences.

The work presented in this thesis looks to unite molecular diagnostics and computational biology techniques and concepts, to contribute to the timely identification, ongoing surveillance and improvement of diagnostic technology, through the development of an automated bioinformatics tool for multi-omics data mining and analysis, as well as a comprehensive, centralized, and expandable oligonucleotide database.

1.1. Highly infectious viruses: a global health perspective

As stated previously, highly infectious viruses pose a persistent threat to global health. According to the World Health Organization (WHO), infectious diseases remain among the top ten causes of death globally, and viral outbreaks continue to destabilize entire regions (World Health Organization, 2024). Beyond increasing morbidity and mortality rates both directly (through infection) and indirectly (by preventing access to treatment for other conditions), outbreaks place an enormous burden on healthcare infrastructures, disrupting already fragile systems in low-resource settings and straining even the most advanced facilities in high-income countries (Haileamlak, 2021).

Moreover, this impact extends beyond the scope of human health, affecting every aspect of society, disrupting infrastructures and threatening the economy. For instance, the 2014–2016 West African Ebola epidemic was estimated to cause the loss of 2.2 billion dollars in the gross domestic product (GDP)

of the most affected countries – Guinea, Liberia and Sierra Leone (Centers for Disease Control and Prevention, 2016a). Similarly, the global cost of HIV/AIDS, including treatment, prevention programs, and lost productivity, has been immeasurable over four decades. To put this into perspective, a 2025 study conducted in the United States concluded that an HIV infection can incur additional lifetime costs of up to 2.4 million dollars per person (Cohen et al., 2025). And most recently, COVID-19 caused unprecedented global economic disruption, with losses running into the trillions of dollars (Cutler & Summers, 2020).

On a societal level, outbreaks disrupt education, commerce, and mobility, while also generating widespread psychological and cultural impacts. The rapid spread of these pathogens often overwhelms health systems and, as seen with COVID-19, has the potential to completely halt travel, trade and education, with the aftermath spanning several years or even decades (*The COVID Decade*, 2021; World Health Organization et al., 2020).

The need for accurate and rapid diagnostic methods should therefore expand beyond acute outbreak responses and persist throughout stable times as well. Preparedness relies on the continuous improvement of current detection technologies and implementation of reliable monitoring systems, to maximize the ability for quick adaptation to any newly emergent or re-emergent threats. The constant evolution of viruses through mutation and recombination requires that diagnostic tools be continuously updated and validated. Without this ongoing effort, even the best-established assays risk becoming obsolete in the face of new variants. Therefore, the creation of robust, sustainable systems for diagnostic design and validation is a public health imperative.

1.2. The Viral Big Three: HIV, Ebola and SARS-CoV-2

1.2.1. HIV

Human Immunodeficiency Virus (HIV) was first recognized in 1981, marking the beginning of one of the most devastating pandemics of modern times. Over 40 million lives have been lost to HIV/AIDS thus far, and another 40 million people currently live with HIV worldwide (World Health Organization, 2025a). While antiretroviral therapies have transformed HIV into a manageable chronic condition in many regions, the absence of a vaccine or cure means that the virus continues to spread, particularly in low- and middle-income countries, as reported by WHO as recently as March 2025.

HIV poses unique challenges for diagnostics. Its rapid mutation rate, driven by the error-prone reverse transcriptase enzyme, results in extraordinary genetic diversity across clades and subtypes (Roberts et al., 1988; Yeo et al., 2020a). This diversity complicates the design of primers and probes, as regions of the viral genome that are conserved in one subtype may vary significantly in another. Furthermore, HIV integrates into the host genome, enabling latent reservoirs that led to the need for developing new assays, including PCR based ones, to be able to achieve reliable detection (Chen et al., 2022). These characteristics make HIV a perfect example of how viral variability can hinder reliable diagnostics, emphasizing the need for continuously updated molecular tools as new variants and adaptative responses emerge.

1.2.2. Ebola

Ebola virus, a member of the Filoviridae family, is responsible for recurring outbreaks, with case fatality rates that can exceed 50% (World Health Organization, 2025b). The 2014–2016 outbreak in West Africa was unprecedented, with more than 28,000 cases and 11,000 deaths, marking the first time Ebola spread across multiple countries on such a scale (World Health Organization, n.d.). This epidemic emphasized both the limitations of diagnostic infrastructure in low-resource settings, and the speed with which viral diseases can overwhelm public health systems.

Unlike HIV, Ebola outbreaks are sporadic, often linked to zoonotic spillovers from animal reservoirs (Alexander et al., 2015). The unpredictable nature of these events makes it essential to maintain diagnostic preparedness even in the absence of active cases. The high fatality rate means that even small delays in diagnosis can have devastating consequences. Molecular diagnostics, particularly PCR-based methods, played a pivotal role during the West African outbreak, but their deployment was often hampered by logistical challenges, including limited laboratory capacity and the need for specialized reagents (Bettini et al., 2023).

1.2.3. SARS-CoV-2 (COVID-19)

The emergence of SARS-CoV-2 in late 2019 led to perhaps the most significant pandemic in over a century. With more than 770 million confirmed cases and over 7 million recorded deaths worldwide

(Mathieu et al., 2020), the scale of COVID-19 was unprecedented. Beyond the health impact, the pandemic caused global disruptions to economies, education systems, and international mobility.

Diagnostics were central to the COVID-19 response. RT-PCR became the gold standard for SARS-CoV-2 detection within weeks of the virus being sequenced, underscoring the value of molecular methods in pandemic preparedness. However, it also exposed critical weaknesses: many regions faced severe shortages of validated primers and probes, quality control was inconsistent, and the rapid emergence of variants of concern raised ongoing questions about assay sensitivity (Ossola, 2020; Public Health England, 2020; Temple-Raston, 2020). COVID-19 demonstrated the urgent need for automated, scalable systems capable of generating and validating primer/probe sets proactively, rather than reactively.

1.2.4. Key similarities and differences

Although they belong to different viral families, these three viruses present key similarities that perfectly illustrate how an analysis tool like the one developed in this project can be engineered to capitalize on them, while still preserving enough differences to show how wide the range of cases to which ViruScope can be applied truly is.

1.2.4.1. Overview and genome organization

The HIV-1 and HIV-2 viruses, causing agents of acquired immunodeficiency syndrome (AIDS), are lentiviruses inserted in the Retroviridae family. This genus is characterized by their ability to cause lifelong infections and high mutation rates, with the differentiating feature from other retroviruses being their tropism for both dividing and non-dividing cells (Carter & Shieh, 2015; *Lentivirus Fact Sheet*, n.d.). HIV is an enveloped, positive-sense single-stranded RNA (+ssRNA) virus, carrying two copies of its genome, also referred to as dimeric RNA (Duchon & Hu, 2024; Moore & Hu, 2009). Being an RNA virus, HIV relies on reverse transcription to produce double-stranded viral DNA, which is then integrated into the host cell's own DNA (Craigie & Bushman, 2012). The HIV genome is comprised of several genes, including Gag (responsible for encoding capsid proteins), Pol (encodes viral enzymes, such as the reverse transcriptase, integrase and protease), Env (encodes the precursor of the mature virion's glycoproteins), and several regulatory regions (Humans, 2012).

Ebolavirus is a genus of the Filoviridae family, made up of six species: *Zaire ebolavirus*, *Bundibugyo ebolavirus, Sudan ebolavirus, Taï Forest ebolavirus, Reston ebolavirus* and *Bombali ebolavirus*, of which only the first four are known to be pathogenic to humans and responsible for the Ebola Virus Disease/Ebola Hemorrhagic Fever (EVD/EHF) (Hoenen et al., 2012). Ebolaviruses are enveloped, negative-sense single-stranded RNA (-ssRNA) viruses, with a characteristic seven gene structure: Nucleoprotein (NP), Viral Protein 35 (VP35), Viral Protein 40 (VP40), GP (Glycoprotein), Viral Protein 30 (VP30), Viral Protein 24 (VP24) and L (RNA-dependent RNA Polymerase), with intergenic regions responsible for regulatory signals (Jun et al., 2015).

SARS-CoV-2, colloquially referred to by the disease it causes, COVID-19, is an enveloped +ssRNA virus belonging to the Coronaviridae family, which has some of the largest genome sizes amongst RNA viruses (Woo et al., 2010). Its genome encodes for 16 non-structural proteins, in which proteins essential for replication are included, such as the RNA-dependent RNA Polymerase (nsp12) and a proof-reading exoribonuclease (nsp14), as well as structural proteins Spike (S), Envelope (E), Membrane (M) and Nucleocapsid (N), which share a high degree of similarity with the proteins of SARS-CoV (Naqvi et al., 2020; Zhao et al., 2021).

These viruses are all enveloped, ssRNA viruses of zoonotic origin (Akoi Boré et al., 2024; Hao et al., 2022; Sharp & Hahn, 2011), with extremely similar entry mechanisms involving membrane fusion and endocytosis processes (Jackson et al., 2022). This is due to the common nature of their glycoproteins – as Class I fusion proteins, they all start out with a precursor that is then cleaved into two subunits, one extracellular and one transmembrane, forming an irreversibly primed homotrimer structure (Kielian & Rey, 2006). Perhaps this is why it comes as no surprise that naturally occurring mutations tend to concentrate in the regions encoding these proteins, and that these variants come accompanied by increased virulence and the ability to better escape immune system detection (Arrildt et al., 2012; Shahhosseini et al., 2021; Wong et al., 2018).

1.2.4.2. Mutations, recombination and intraspecies conservation

Among the three, HIV exhibits the highest mutation rates and variability, as a consequence of its error-prone reverse transcriptase – which lacks proof-reading ability –, host Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3 (APOBEC3) editing, frequent recombination, and its life-long infection period that allows for in-host evolution (Rhodes et al., 2005; Yeo et al., 2020b; Yu et al., 2018).

In vivo estimates reach ~4×10^{-3} mutations/base per cell in intracellular complementary DNA (cDNA) (dominated by APOBEC3 editing), while per-replication reverse transcription (RT) error rates are typically lower (10^{-5}) but still considered high by cellular standards (Cuevas et al., 2015). This results in a global landscape of subtypes and HIV CRFs (Circulating Recombinant Forms) that can differ substantially at primer/probe sites.

Ebolavirus, on the other hand, evolves more slowly than HIV, but still accumulates substitutions across and within outbreaks; lineages can diverge in severe ways, with the most impactful mutation occurring in the protein encoded by GP during the 2013–2016 outbreak, where the alanine to valine amino acid substitution at position 82 (A82V) resulted in higher levels of infectivity and mortality (Diehl et al., 2016). However, its shorter chains of transmission due to acute disease limited within-host diversification when compared with chronic HIV.

Interestingly, SARS-CoV-2 mutates slower than typical RNA viruses. This is because, unlike the previous two viruses, it possesses proof-reading ability. Nsp14 proofreading elevates replication accuracy (Denison et al., 2011; Gribble et al., 2021); nevertheless, the pandemic enabled rapid rates of mutations and recurrent recombination due to its unprecedented scale, culminating in the appearance of variants of concern, all of which had their mutations heavily concentrated in the spike protein region, the same region primarily targeted by vaccines and detection methods (Hendy et al., 2021).

1.2.4.3. Implications for detection and diagnostics

In terms of the effect the aforementioned characteristics can have on the ability to design and upkeep reliable detection and diagnostic methods, we must consider both the convergence and divergence points between these viruses and what they translate to on a practical level.

Firstly, we can conclude that RNA genomes are inherently unstable over time: prone to replication errors, reliant on recombination and under constant evolutive pressure → primers and probes must primarily target the most highly conserved regions to avoid false negatives.

Secondly, enveloped viruses are easily targetable due to their surface proteins, and their inhibition can lead to complete viral inactivation. On the flip side, being a major antigenic site naturally promotes

evolution, which is accelerated in viruses already → assay sensitivity can be disrupted in short amounts of time if the diagnostic target overlaps with variable regions in the sequences.
Thirdly, intra-species variability is present in all cases, threatening the ability of the primers to detect significant amounts of variants → diagnostic pipelines must be designed using consensus-based strategies and should support the use of degenerate primers.

Finally, constant monitoring of the population is key for identifying the current circulating strains and updating assays. Additionally, the zoonotic origin of viruses like the above indicates that surveillance of animal reservoirs and the development of protocols for testing and monitoring potential spillover zones should also be taken into consideration.

1.3. Molecular diagnostics and the role of oligonucleotide databases in viral detection

1.3.1. PCR: The gold standard for detection

Polymerase Chain Reaction (PCR) remains the benchmark for viral diagnostics due to its unique combination of sensitivity, specificity, and rapid turnaround time. PCR can detect minute quantities of viral nucleic acid, making it invaluable for early-stage infections when viral loads are low (Garibyan & Avashia, 2013). Its high specificity derives from carefully designed primers and probes, which bind to target regions of the viral genome. Moreover, PCR assays can be rapidly developed once viral sequences are available, as demonstrated by the emergency release of newly developed Ebola PCR kits during the 2014-2016 outbreak by the CDC (Centers for Disease Control and Prevention, 2016b, 2016c) , and more recently during the COVID-19 pandemic (Madadelahi et al., 2024).

However, PCR's reliability depends fundamentally on the design of its oligonucleotides. Poorly designed primers may bind non-specifically, form dimers, or fail to amplify across different viral strains (Bustin & Huggett, 2017). Consequently, databases that curate validated oligonucleotides play a critical role in maintaining assay quality and reproducibility. Without centralized and continuously updated repositories, researchers are forced to search across scattered literature or continuously design their own primers, increasing the risk of redundant or suboptimal assay designs, as well as being time consuming.

1.3.2. HIVoligoDB, EbolaID and CoV2ID

Several virus-specific databases have been created to consolidate oligonucleotide information. HIVoligoDB, EbolaID, and CoV2ID (J. Carneiro et al., 2017, 2020; J. Carneiro & Pereira, 2016a) are among the most notable examples, providing researchers with centralized repositories of primers and probes validated for their respective viruses. These resources were valuable for their time: they made data openly accessible, reduced the fragmentation of information across disparate sources, and facilitated reproducibility. Yet, their limitations are equally clear. Each relied on manual curation, a process that is inherently slow and difficult to scale. As a result, updates to these databases were irregular, and many have since become inactive. For instance, databases such as VirOligo (Onodera & Melcher, 2002) and the Ebola Database (EBD) (Swetha et al., 2016) are no longer online. Their virus-specific focus also limited broader applicability: while specialized repositories can address immediate needs, they do not support cross-viral comparisons or rapid adaptation to novel pathogens. The limitations of these databases highlight the urgent need for automation, scalability, and cross-viral scope in future resources. Furthermore, tools that can be quickly deployed to analyze huge amounts of genomic data in case of a sudden emergence, process it for possible primers/probes, and have it ready for upload into a centralized database for global access within a short timeframe, can drastically cut down on the time needed to develop a reliable detection assay, as researchers would not need to start the curation process from scratch.

1.3.3. Challenges and research gap

At present, oligonucleotide data remains scattered across thousands of individual research papers, making it challenging for scientists to identify, validate, and reuse existing information. Manual curation, while valuable, cannot keep pace with the volume of data being generated, particularly for highly studied viruses such as HIV, Ebola, and SARS-CoV-2. This reliance on human effort is both time-consuming and error prone.

More critically, there is no generalized pipeline capable of automatically retrieving, generating, and validating primers across multiple viral genomes. Current approaches are fragmented, virus-specific, and dependent on manual intervention. This gap limits the speed of diagnostic development and weakens global preparedness. The absence of automation is particularly problematic in the context of emerging threats, where time is critical and delays can cost lives. Given this, a tool capable of automated sequence

mining, primer generation, validation, and scoring, coupled with a continuously updated cross-viral database, would represent a transformative advance in molecular diagnostics.

1.3.4. Hypothesis

Considering the above contextualization, we hypothesize that the integration of viral genomic data and literature-derived oligonucleotide information into a single automated bioinformatics platform will streamline the discovery, design, and validation of primers for diagnostic assay development. By automating processes such as sequence mining, primer generation, and alignment-based validation, the tool and database are expected to overcome the scalability issues of current approaches and fragmentation of viral information, respectively.

This hypothesis frames the central research question of the thesis: Can automation and integration across data sources enhance the efficiency, reliability, and cross-viral applicability of primer design for diagnostic purposes?

1.4. Research aim and objectives

The primary aim of this project is to develop and deploy a bioinformatics tool capable of biological data extraction and analysis, designed to overcome the limitations of existing resources by integrating automation, scalability and cross-viral applicability. Additionally, an open-access database created using the tool's functionality will be made available as proof of concept and as a centralized information hub for viral diagnostics.

This project is designed to equip researchers and technicians with a comprehensive suite of tools that streamline and automate key processes in viral diagnostic assay development. The primary objectives are as follows:

- Automated Mining and Analysis of Viral Sequences: The tool will enable users to automatically retrieve and analyze viral genomic sequences, facilitating efficient and thorough examination of relevant data.
- Text Mining of Literature for Oligonucleotide Extraction: By leveraging automated text mining techniques, the tool will extract validated oligonucleotide sequences from scientific literature, consolidating valuable experimental information.

- In Silico Generation of Novel Primers: The system will support the computational design of new primers, expanding the pool of candidates available for PCR-based detection of viruses.
- Validation Against Genomic Alignments: All candidate primers, both extracted and newly generated, will be validated through comparison with multiple sequence alignments, ensuring their specificity and effectiveness across viral strains.
- Automated Primer Pair Combination: The tool will automate the pairing of compatible primers, streamlining the creation of primer sets optimized for diagnostic assays.
- Calculation and Scoring of Benchmark Primer Parameters: Each primer and primer pair will be evaluated on critical parameters, including GC content, melting temperature, conservation across strains, and stability against secondary structure formation. These benchmarks ensure the reliability and performance of the primers.
- Export of Structured, Clean Outputs: The final outputs will be generated in a structured and user-friendly format, ready for deployment in downstream applications or integration into centralized databases.

Together, these objectives aim to create an integrated pipeline that enhances the speed, accuracy, and reproducibility of viral diagnostic assay design, providing the scientific community with a robust resource for present and emerging viral threats.

The database creation, on the other hand, looks to merge and expand previously existing databases (HIVoligoDB, EbolaID, CoV2ID) using automated processes, establishing a FAIR (Findable, Accessible, Interoperable, Reusable) cross-viral repository (Wilkinson et al., 2016).

The combination of these two major objectives addresses a critical gap in molecular diagnostics left by the disappearance or lack of maintenance of previous databases, providing researchers with up-to-date primer data to potentially help accelerate the response capacity for emerging viral threats. The conservation aspect of primer analysis and design also looks to facilitate detection across viral strains, improving surveillance capacity. Furthermore, the establishment of an adaptable workflow that can be deployed for other viruses makes ViruScope, the tool, and ViruScopeDB, the database, able to deliver long-term, expandable and speedy resources for data integration from both past and future outbreaks.

In sum, this work aims to accelerate the development of reliable detection assays and contribute to global preparedness against the threat of infectious viruses.

2. Methodology

The project's workflow can be divided into three main sections, according to the type of data being processed: analysis of complete genomes, scraping, generation and scoring of oligonucleotide datasets, and, finally, compilation of outputs into tools, a software application and a database (**Figure 1.**)

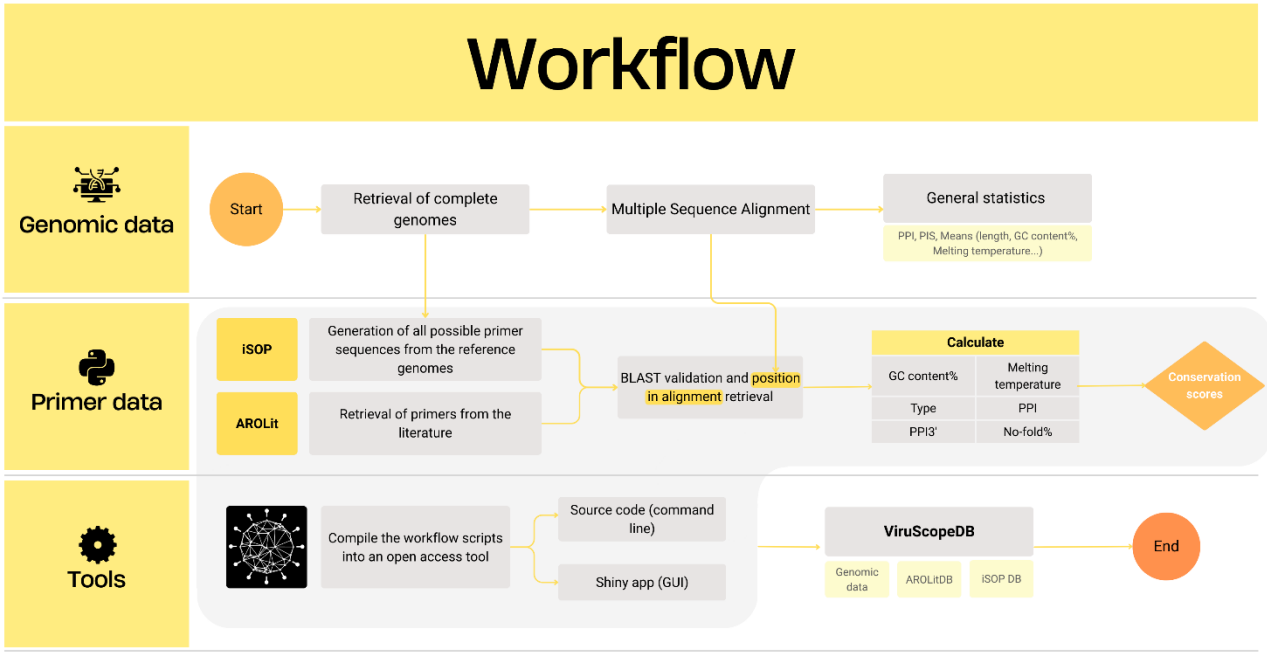


Figure 1. Workflow of the entire project divided according to the type of data being processed. First row illustrates the process of collection and handling of data pertaining to the genomes of each virus. Second row shows how primer data was obtained through both literature scraping and *in silico* generation, and the subsequent validation and analysis of the outputs from each tool. The last row describes the rework of the scripts developed for the previous steps into a ready-to-use tool, made available both as a command line program and as a full application with a Graphical User Interface (GUI). The final step consists of uploading the outputs of the ViruScope tool to an online database.

The first step (Figure 1, row 1) encompasses the collection, processing and analysis of all complete genomic data for Ebolavirus, HIV and SARS-CoV-2 (reference sequences and all strains). After sequence collection, a Multiple Sequence Alignment (MSA) was performed on each set, and information was extracted from the outputs, including – but not limited to – the percentage of pairwise identity (PPI), proportion of identical sites (PIS), mean length of the genomes, GC content percentage, coding DNA sequences (CDS), and more.

The complete reference genome for each virus was then used in the second phase (Figure 1, row 2), to generate primer data *in silico*. The literature that detailed the detection of these viruses using PCR was retrieved from PubMed and scraped for experimental primer data. The genomic data from the previous step was used in the validation process of both primer datasets, and the MSA output served to determine where in the alignment each primer would bind to the DNA strand. The primers, as well as the primer pairs formed, were then scored and ranked based on parameters crucial to the success of a PCR run, such as

GC content (%), Melting Temperature (Tm, °C), Type (Forward or Reverse, PPI (%), percentage of pairwise identity at the 3' end (PPI3', %), and probability of not forming unwanted secondary structures (No-fold, %), constituting the final Conservation Score.

Finally, the last section (Figure 1, row 3) consisted of preparing the custom scripts used to obtain the previous step's results for general use, originating the ViruScope tool, available for both the command line and as an application with a graphical interface for ease of use. The outputs of the tool were then compiled and made available in the form of a website, ViruScopeDB, to serve both as a publicly accessible database and proof of concept of the tool.

All analyses were performed on a personal workstation running Microsoft Windows 11 Home (Version 10.0.26100, 64-bit) on an HP Victus 16-e0xxx laptop, equipped with an AMD Ryzen 5 5600H processor (6 cores, 12 threads, 3.3 GHz), 8 GB DDR4 RAM, and 512 GB SSD as primary storage. The software environment included Python 3.12.3, run in Visual Studio Code 1.104.2, and the Windows Subsystem for Linux (WSL) running Ubuntu 24.04.3 LTS (Noble).

2.1. Viral multi-omics data mining

The first step in the project pipeline was the collection of all the available and validated complete genomic sequences for each virus. Public databases were extensively searched and filtered for this purpose, and the collected sequences were subsequently aligned, analyzed and used in the AROLit and iSOP workflows for validation and information extraction for the primer sequences.

2.1.1. Genomic data filtering and collection

Genomic sequences for each virus were obtained from the National Center for Biotechnology Information (NCBI)'s dedicated viral database, NCBI Virus (accessed October 24th, 2024). The "Nucleotide completeness" filter was applied with the objective of only collecting complete genomes for export. Further refinement of search parameters included the taxonomic ID of each major strain group of interest (e.g.: all ebolavirus species capable of causing EVD in humans), to ensure the capture of all sequences belonging to the main virus groups (HIV, Ebolavirus, SARS-CoV-2). For SARS-CoV-2 specifically, extra filters, "Max Ambiguous Characters" and "Minimum/Maximum Length" were added to reduce the number of sequences to a number that could be analyzed with the computational power available, while still maintaining a representative sample size.

Data processing involved reordering the FASTA file output so the reference sequence (RefSeq) appeared at the top, avoiding its elimination in the subsequent duplicate removal step. Both steps were performed using custom scripts made available under "utilities_fasta.py" and the Utilities tab in the final application.

2.1.2. Multiple Sequence Alignment (MSA)

The collected sequences were then uploaded to Galaxy (last accessed 11th of March 2025) to perform the alignments. The algorithm used to perform the MSA was MAFFT (Multiple Alignment using Fast Fourier Transform) 7.526, due to its high accuracy even in sequences that have extensions or are more distantly related. Furthermore, the chosen heuristic, the progressive method FFT-NS-2, allows for accuracies comparable to other MSA algorithms, while drastically reducing CPU time, making it ideal for the volume of sequences being analyzed (Katoh et al., 2002).

2.1.3. Gene extraction and mapping to alignment

Automatic retrieval of the loci for the genes of each virus was done using a custom script that iterated over a Genbank file and extracted any features identified as "Gene" or "CDS". Loci that appeared multiple times with different start and end positions were merged into a single entry, matching the positions reported by NCBI Nucleotide's Graph View of the RefSeq entries (**Figure 2**).

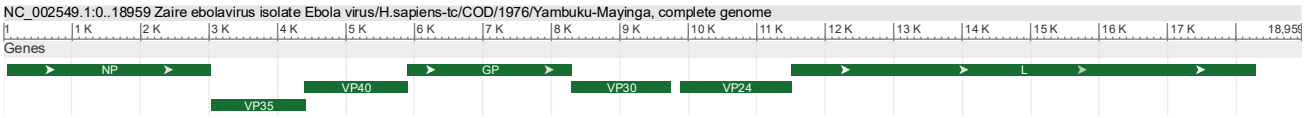


Figure 2. Gene map of *Zaire ebolavirus* isolate Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga (NC_002549.1). Hovering over each gene reveals its starting and ending positions. (Zaire Ebolavirus Isolate Ebola Virus/H.Sapiens-Tc/COD/1976/Yambuku-Mayinga, Complete Genome, *2018)*. Retrieved on September 15th, 2025 from NCBI Nucleotide Database, U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/nuccore/NC_002549.1?report=graph#.

To map the gene coordinates in the reference genome (ungapped) to the one in the alignment file (containing gaps (-) introduced by MSA), the script proceeds to iterate over the alignment sequence, and, each time a non-gap character is encountered, the reference index is recorded as mapping to the alignment index. For each locus, the original coordinates were projected onto the alignment by converting them to alignment coordinates. The script also takes into account the difference between Genbank and Python indexing, with the first being 1-based and the second 0-based, adjusting the final positions

accordingly. If both start and end positions are successfully mapped, the output returns the list of genes with coordinates specific to the input alignment.

Finally, the resulting list of coordinates was compared to the ones obtained from Geneious Prime's "Annotate from Database" feature, available in the premium version, to validate the output from the function.

Once the primer datasets were ready, each primer had a locus assigned to it based on their positions in the alignment. Primers with positions not matching any loci, or only partially in a locus, were classified as "Intergenic/Unknown".

2.2. AROLit v.3: tool development and optimization

Automatic Retrieval of Oligonucleotides from Literature (AROLit) is a tool first developed by Carneiro et al. in 2023, and later further optimized by Pinheiro et al. (2024), with the goal of automatizing the retrieval of oligonucleotide sequences, namely primers, directly from articles, with minimal user interaction. Initially, the established pipeline was going to be reworked only with the intent of generalizing its use, given that a significant portion of the scripts was hardcoded for the organisms relevant to the projects at the time. However, AROLit relied on external software for both the PDF-to-text conversion and primer-scraping steps, making it necessary to do a complete overhaul of the script due to two major problems: PDF Shaper, the tool that handled the conversion of the articles to the easily-parsed .txt format, changed its business model to only allow conversion of more than five PDFs if the user had a paid subscription. Secondly, the tool that ran the scraping algorithm, Orange Data Mining, failed repeatedly in the very first step of the workflow – loading the .txt files – and even when the attempt was made to submit a test file directly as a text box input, the workflow failed again on the first Regular Expression (RegEx) block, which simply recoded all new lines ("\n") to null characters ("").

These setbacks meant that a completely new workflow had to be implemented, although the logic behind it would stay the same. Furthermore, run time optimizations, scalability, generalization of the original code and minimization of workflow interruptions ("hopping" between different programs as opposed to having a single, continuous pipeline) were all objectives of this section of the project, giving way to the development of the third version of AROLit.

2.2.1. Article retrieval

To obtain open-access articles with experimentally validated primer sequences, PubMed Central was queried using the Entrez module for Unix-based command lines (November 20th, 2024 for Ebola and HIV; December 4th, 2024 for SARS-CoV-2), allowing for the bypass of the BioPython version of the same module and the website's limitation – retrieval of the first 10,000 results only. For each virus, the search query followed PubMed's syntax - "'[Virus name]' AND 'PCR'". The output was saved in Medline format to a .nbib file, a citation file format ready for upload to citation management softwares.

Zotero v. 7.0 was chosen as the reference manager due to two crucial features for the article collection process: the ability to automatically retrieve the full-text PDF (for the articles where it is available), and the ability to export the PDFs using the digital object identifier (DOI), as the file name (excluding the "/" character due to file naming conventions), allowing for the maintenance of the "Primer ‹–› Article" association throughout the entirety of the workflow.

2.2.2. PDF to text conversion and primer scraping methods

To replace PDF Shaper and Orange Data Mining for the PDF to .txt conversion and primer scraping steps, respectively, several methods were tested to determine which one presented with the highest accuracy. Other important aspects of the replacement taken into consideration were its dependency, or lack thereof, on external software, and the hardware requirements needed to run the tool.

Each algorithm was run against a test batch consisting of 21 random HIV-related articles that contained primers, either in the inline text or in tables/figures, and 10 PDFs with no primers, to serve as the negative control. The primers specific to HIV were manually retrieved and inserted into an Excel table, amounting to 119 primers in total. To automatically cross-check the test primers with the different methodology's outputs, a formula was coded to check if the value of each row (the primer sequence) could be found, in its entirety, anywhere in the list of primers that were automatically extracted, meaning that partial matches were considered negative results. The results column for the method being evaluated then populated itself according to one of two conditional cell formattings: green fill/"YES" for positive identifications, and red fill/"NO" for sequences that were not successfully extracted.

2.2.3. Method testing and evaluation

Several different methods were tested to determine which could detect the most amount of test batch primers. The designed approaches can be divided into two major categories: Large Language Model (LLM)-based and Classic.

LLM models were the first to be considered as a replacement for the obsolete pipeline, given their spike in popularity and widespread use in recent years, a consequence of the exponential growth of their capabilities. These models can interpret context, potentially yielding accurate results with less data processing than required for a uniform primer database. For instance, a well-trained model could extract specific virus sequences from PDFs, even when primers for multiple viruses appear in the same article. Furthermore, the possibility of having the outputs already formatted for direct CSV or SQL inputs was also a vantage point for this approach. On the flip side, open-access LLM models are either computationally expensive to run and host locally, which is a major disadvantage, especially considering the end goal of building an open-access tool, or do not have enough parameters to be able to accurately retrieve the data – both things complicated further by the amount of data being analyzed. Conversely, application programming interface (API) keys, or running data analysis directly on a model's website, has costs associated. OpenAI's ChatGPT 4, the best model available at the time of testing, managed to perfectly retrieve the primers in an article – flagged in the test batch as having complicated formatting – and prepare them so they could be directly made into a CSV file. However, the free plan had, and continues to have, heavy restrictions on data extraction and analysis prompts, usually only allowing the parsing of single digit amounts of documents at a time. To circumvent the restrictions, we would have to run the model on a pay-per-token basis, which is unfeasible for this project.

Another model tested was Google's NotebookLM, which showed promising results at first. However, two major problems were quickly identified; First, the maximum amount of PDFs inputs per chat was limited to 50. This is not an obstacle for a number of articles around the size of Ebola's set, 256 PDFs, requiring the creation of only 6 chats. This input method, however, does not scale well for datasets two or more orders of magnitude bigger, as is the case with the HIV and SARS-CoV-2 articles. Other disadvantages of this approach included the sharp drop in accuracy as the number and complexity of articles increased, the lack of reproducibility of the results (different chats yielded different primers, even with the same prompt, was prone to hallucinations), failure at creating the "Primer ‹–› Article" link in its entirety, and its inability to be seamlessly integrated into the final pipeline.

The final LLM-based approach tested used HuggingFace's Python module, unlocking the possibility of running models locally or making free API requests to their servers. Models like MinerU 2.5 demonstrated an incredibly reliable extraction of PDF content to Markdown format, even in articles where all the other converters failed to parse. Unfortunately, to run the model locally, the hardware available did not meet the requirements as per the developers' GitHub page. While HuggingFace allows the user to send requests to a model hosted on their servers, the free tier is only useful for light testing, making it unviable as a converter of thousands of multi-page documents, and MinerU was not one of the models available in the paid tier either way. The independent application only allows for the conversion of 20 PDFs at a time, making it unscalable too. A pipeline was built using smaller models, DonutAI for document conversion and OpenLLaMa-7b for extracting information given a detailed prompt, however, a test run done with a single file quickly revealed that 1) the conversion was not entirely successful and 2) when provided with a simplified test paragraph containing one primer sequence, the returned output not only took over 300 seconds, but the identification of the sequence was incorrect.

Given the string of failures surrounding the LLM approach, we took a step back and tested methods that made use of more classic, well-established tools. For the first half of the pipeline, the PDF conversion, the search started with trying to find software that could replace PDF Shaper. ABBYY FineReader was selected as the candidate. An optical character recognition (OCR) approach, similar to what web browsers use to make some image-only PDFs highlightable, was also prepared for comparison, using Tesseract/PyTesseract. The final conversion tool tested was a Python module specifically designed for this task, PyMuPDF, later replaced with PyMuPDF4LLM due to having more features and performing better. For primer extraction of the resulting outputs, we used RegEx, as primer sequences are distinct enough that context is not necessary. Additionally, manual analysis of cases where the extraction failed revealed that the problem lay in the conversion step and not the extraction, with methods failing to convert some structures, such as tables or images, that were more complex, completely losing the primer data of the article. The same extraction pattern (described in more detail in **Section 2.2.4**) was applied for all converters.

Tesseract, the OCR approach, performed the worst, being discarded immediately. ABBY FineReader slightly underperformed in comparison with PyMuPDF4LLM. Combined with the fact that it also required a switch to an external software with no Python integration, unlike PyMuPDF, made it so the chosen

method for the new AROLit pipeline was a combination of PyMuPDF4LLM (conversion step) + RegEx (data extraction step). Further attempts to improve accuracy by adding fallbacks to the conversion step (e.g.: extracting tables that otherwise would not make it into the final PyMuPDF4LLM output as images, which the module does reliably, to be analyzed by Tesseract in a complementary step) did not yield better results, making the pipeline structure final.

2.2.4. Primer retrieval using Regular Expressions (RegEx)

Once established that the combination of PyMuPDF4LLM+RegEx was the most cost effective method for primer retrieval, both in regards to its low computational power requirements and relative speed, as well as being lightweight enough to run on most machines, the algorithm was further refined in order to capture the most amount of primers, without compromising accuracy by making it too complex and overfitted to fringe cases where primers are reported with unconventional formats.

The choice to use RegEx over more modern methodologies quickly proved itself to be beneficial once again, as its short run time allowed for the implementation of two different RegEx iterations: a generalized one, capable of capturing most primers, and a specialized one, capable of detecting primers with more complex or unusual formatting, at the cost of losing out on a portion of the more simple ones (**Figure 3**). This combination further boosted the accuracy of the algorithm, while only adding an imperceptible increase in run time.

```
FUNCTION generalized_RegEx(article):
    pattern ← regex that matches:
        - NOT preceded by a letter OR digit
        - ≥ 17 bases from set {A,C,G,T,U,R,K,S,M,Y,W,B,H,N,D,V} possibly separated by spaces OR dashes
        - NOT followed by a letter OR digit

    DEFINE clean(sequence):
        REMOVE spaces OR dashes from sequence
        Convert to uppercase

    hits ← []
    FOR each match in article USING pattern:
        CALL clean(match)
        Append match to hits
    END FOR

    seen ← {}
    hits ← [
      IF the length of the sequence is ≥ 17 AND sequence NOT in seen THEN
            ADD sequence to seen
      ELSE
            REMOVE sequence from hits
            ADD sequence to seen
      END IF
      ]

    RETURN hits
```

```
FUNCTION specialized_RegEx(article):
    pattern ← regex that matches:
        - NOT preceded by a letter OR digit
        - sequences of symbols from set {A,C,G,T,U,R,K,S,M,Y,W,B,H,N,D,V, /, (, )} WITH optional spaces OR dashes
        - NOT followed by a letter OR digit

    DEFINE clean(sequence):
        REMOVE spaces OR dashes from sequence
        Convert to uppercase

    hits ← []
    FOR each match in article USING pattern:
        CALL clean(match)
        Append match to hits
    END FOR

    FOR each sequence IN hits:
        REMOVE spaces from h
        STRIP parentheses () from h
    END FOR

    seen ← {}
    hits ← [
      IF the length of the sequence is ≥ 17 AND sequence NOT in seen THEN
            ADD sequence to seen
      ELSE
            REMOVE sequence from hits
            ADD sequence to seen
      END IF
      ]

    RETURN hits
```

Figure 3. Pseudocode illustrating the logic behind the generalized (left) and specialized (right) RegEx functions.

Every primer detected by the generalized RegEx was stored in the final list of sequences, whereas the primers captured by the specialized expression were only added to the final output if they were "new", in an operation akin to a set union.

2.2.5. Primer validation through NCBI BLAST

Following primer extraction from the articles, the sequences were exported to a FASTA file, in order to perform a validation step. Since articles could include experimental data for viruses other than the ones that serve as a focus of this project (e.g.: articles that detail PCR experiments for both HIV and HBV), and given that the scraping pipeline, while accurate, is not foolproof, two BLASTs were conducted using the NCBI BLAST+ command line tool. Two BLAST databases were created, one for the ungapped reference genome and the other for the genome in the alignment, to retrieve the positions of the valid primers in both cases. Both runs used the same parameters, described in **Table 1**.

Table 1. Options used in the "run_blast()" command.

Parameter	Value used	Description
BLAST type	blastn-short	Searches a nucleotide query against a nucleotide sequence or database
gap-open	1	Cost to open a gap
word-size	10	Length of exact initial match
evalue	0.01	Expect value for saving hits

The type of BLAST used was blastn-short, due to its optimization for querying short sequences, such as primers, probes or PCR fragments. The word size parameter was also chosen with the characteristics of the sequences in mind, being smaller than the default value, triggering extension at 10 identical consecutive bases (around half of an average primer), but not too small (like 7-8 nucleotides) as to catch non-relevant sequences that might show up by chance, especially in large genomes. The cost to open a gap was decided based on the rationale that while very short binding sites rarely contain indels, alignments in real genome panels and MSAs can show single-base gaps relative to the reference (true indels or alignment artifacts). Therefore, a non-null but still low gap-open value allows the BLAST to report cases where there may be mismatches as to not lose important information (e.g.: one mismatch,

especially at the 5′ end, is not critical for a PCR experiment), while still ensuring, through the calculation of PPI and PPI3′ downstream in the pipeline, that any potential gaps are correctly addressed. The low e-value, set at 0.1, ensures exact or near exact matches are counted as hits and simultaneously filters out low-significance extensions.

The matches found for each run were then cross-referenced and merged, creating a primer list detailing the hit sequences and their positions within both the ungapped and the aligned genomes.

2.3. iSOP v.3: tool development and optimization

Developed simultaneously with AROLit by Carneiro et al. (2023), the *In Silico* Oligonucleotides designed in Python (iSOP) tool can generate all possible primer sequences of various lengths from a reference genome. The script also allows for the mapping of the output primers to an aligned version of the RefSeq, with subsequent BLAST validation. The functions pertaining to this specific module stayed mostly the same as its predecessor versions, with changes primarily focusing on optimizations with the goals of drastically reducing runtime and making it able to handle datasets several times bigger than the ones the tool had been used to process up until this point.

2.3.1. *In silico* primer generation

To generate all possible primer sequences, a "sliding window" approach was employed, taking as parameters the minimum and maximum primer lengths, and a step value (**Figure 4**).

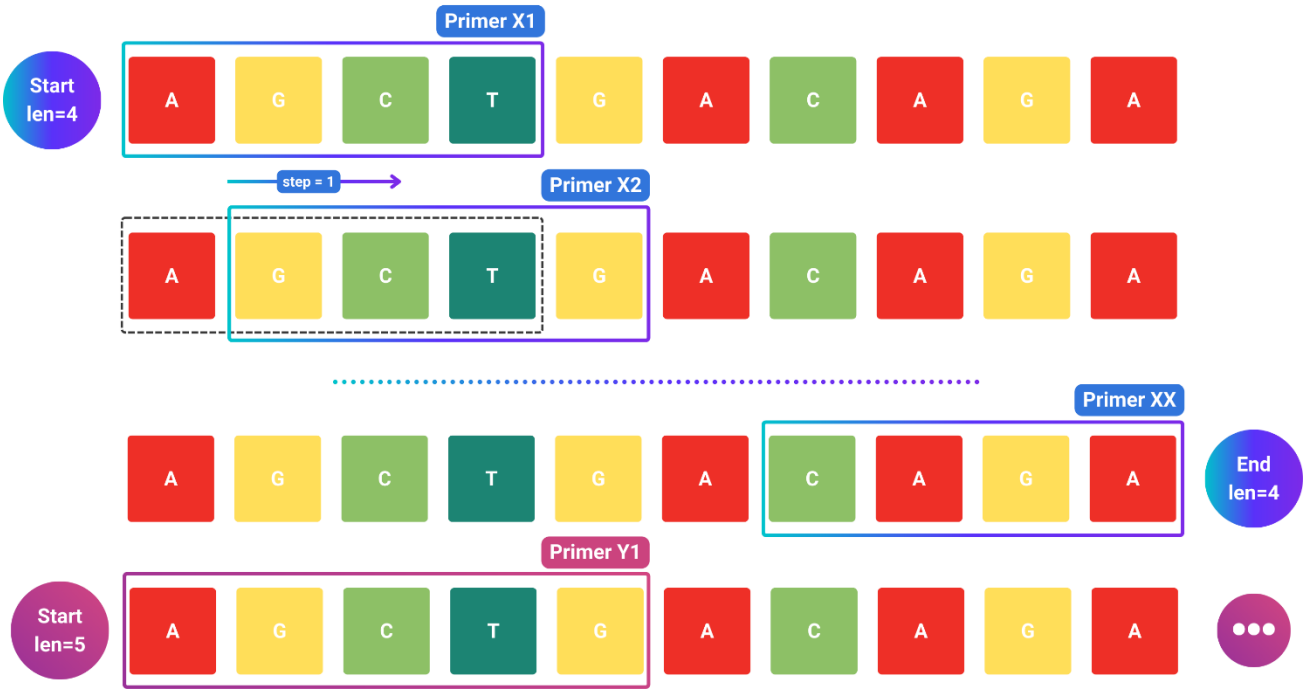


Figure 4. Visual illustration of the sliding window method for a minimum primer length of 4 (len=4), maximum length of 5 (len=5), and a step size of 1. After "sliding" across the entire genome and saving all possible primers with 4 nucleotides, the window returns to the beginning of the sequence and repeats the process to generate all primers with 5 nucleotides.

In tandem with the generation of primer sequences, essential metadata is systematically recorded for each primer. This includes information such as the primer "Type" (indicating whether it is a forward or reverse primer), as well as the start and end positions within the reference genome, and the overall length of the primer. Throughout this process, placeholder values are assigned to any dictionary keys that do not yet have corresponding data. These placeholders ensure that the structure of the data remains consistent for all primers, regardless of when specific values are calculated. This uniformity across all primer datasets facilitates streamlined processing and analysis further along the pipeline.

2.3.2. Validation of the generated primers

As the sequences were generated directly from the genome, the validation step and position retrieval for iSOP-based primers consisted of only the BLAST run against the reference in the alignment, as detailed in **Section 2.2.4**, eliminating sequences created from poorly conserved, heavily gapped regions.

2.4. Primer scoring

After validating both primer datasets, each were individually scored based on the same parameters and logic established and used in previous iterations of the project (F. Carneiro et al., 2023; Pinheiro et al., 2024). The functions were rewritten to have reduced running times and be applicable to any virus (or organism with similar genomic arrangements), by reworking every hardcoded section of code. The

algorithms themselves were also revamped, giving more nuance to the scoring system by introducing experimentally relevant parameters into the calculations performed, aiming to further shorten the gap between computational predictions and experimental data.

2.4.1. GC Content and Melting temperature

GC Content (GC, %) and melting temperature (Tm, °C) calculations stayed largely the same as in previous versions of the tool, with one key difference: the handling of degenerate bases, present in a non-insignificant portion of the primers described in literature. An attempt to rapidly calculate these parameters was made using AltaiR (Silva et al., 2024), however, even with alterations made in the software to include degenerate bases in calculations, the inherent architecture of the software only allowed for randomized mapping of these nucleotides to the base ACGT ones. So, to avoid this, and given that GC and Tm calculations were not computationally taxing, a custom approach was chosen instead.

For this purpose, a function was created, "expand_degenerate()", to generate all possible sequences for a primer containing degenerate bases. From this, two approaches were implemented: one set of functions that calculate the GC content and Tm of all sequences resulting from the previous function, treating them as individual primers and returning them identified as variants of the initial sequence, and another set that returns an interval of values for each parameter, assigned to the original sequence. An illustrative example of how this function works is described in **Figure 5**.

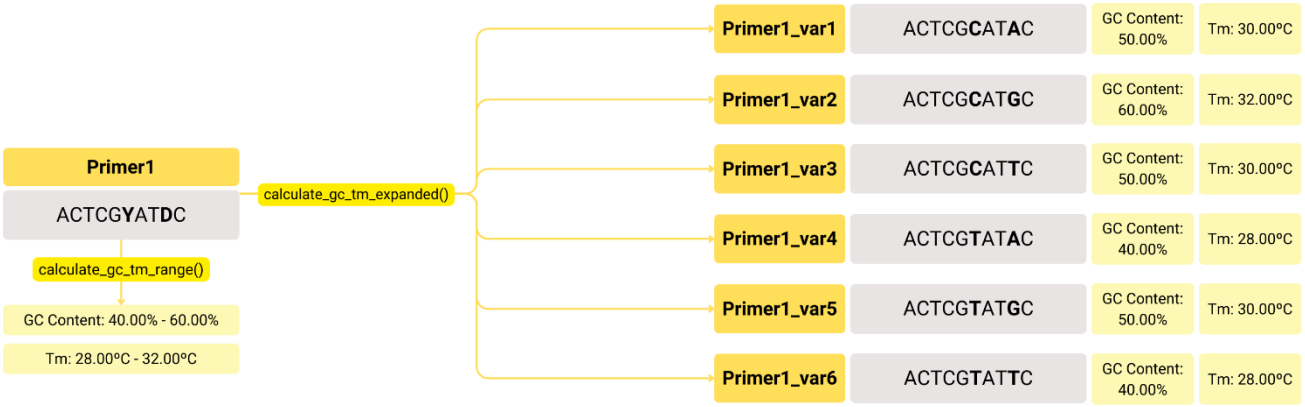


Figure 5. Example of how a primer with two degenerated bases is processed. IUPAC nucleotide code "Y" indicates the actual base at that position can either be a cytosine or a thymine. Similarly, base code "D" can be adenine, guanine or thymine. Thus, by the multiplication principle, the number of possible sequences Primer1 can be is 6 (2 x 3). The "expand_degenerate()" function is nested in both calculate functions declared above.

Both functions provide important information depending on the use case, but for the subsequent calculations ran in the tool, the results from the first function were the ones used.

The formulas for calculating the Tm of each primer were adapted from BioPHP's Melting temperature calculator (https://www.biophp.org/minitools/melting_temperature/), and are as follows:

$$Tm(°C) = \begin{cases} (wA + xT) \times 2 + (yG + zC) \times 4, & n \leq 13 \\ 69.4 + 41 \times \dfrac{yG + zC - 16.4}{wA + xT + yG + zC}, & n > 13 \end{cases} \quad (1)$$

Where *w*, *x*, *y* and *z* represent the number of times that nucleotide shows up in the sequence, and *n* is length of the sequence in nucleotides. This equation assumes that the annealing occurs under standard conditions, as in, when the concentrations for primers and Na^+ are 50 nM and 50 mM, respectively, and pH = 7.0.

The formula used for GC content calculations was:

$$GC\ (\%) = \frac{yG + zC}{wA + xT + yG + zC} \times 100 \quad (2)$$

2.4.2. Conservation scores

The concept of the "Conservation Score" for an oligonucleotide, as defined by Carneiro et al. in 2023, is the arithmetic mean of two parameters: the Percentage of Pairwise Identity (PPI) and the Percentage of Pairwise Identity at the 3' end (PPI3'). This score is crucial for determining whether a given region is sufficiently conserved to serve as an effective binding site for a designed primer. By combining overall sequence similarity (PPI) with the conservation of the critical 3' terminal bases (PPI3'), the Conservation Score provides a comprehensive measure of the likelihood that the primer will successfully anneal to its intended target across different sequence variants.

$$Conservation\ Score\ (\%) = \frac{PPI + PPI3'}{2} \quad (3)$$

To determine the values of these two parameters, an extra function was created to classify the literature primers by "Type", which corresponds to either "Forward" or "Reverse". *In silico* primers were automatically classified during the generation process. This step is not only important for creating primer pairs later, but it is also crucial to the determination of the PPI3' in particular.

First, an ambiguity map was created to account for the literature primer's inclusion of IUPAC degenerated nucleotide codes when calculating sequence similarity. Each entry represents the probability of two nucleotides being considered identical. For example, an adenine (represented by "A") has a 50% chance of identity with the nucleotide in a position that's represented by "R" – this is because that code represents either an adenine, A, or a guanine, G. Thus, the weight attributed to "A" is going to be 0.5, and the same goes for G. Similarly, other ambiguous codes such as "N", which can represent any nucleotide, are attributed equal weights across all four bases (0.25). This allows for a probabilistic approach to handling degenerated bases in PPI calculations, rather than a binary one.

To quantify similarity, first we defined a function named "pairwise_identity()" to process individual columns, where an initial check is performed to automatically discard columns where fewer than 30% of the sequences contain a valid nucleotide. This threshold was set based on Geneious Prime's logic, where pairwise identity is segmented into three categories: green regions (100% identity), green–brown regions (30–100% identity), and red regions (< 30% identity), as illustrated in **Figure 6**.

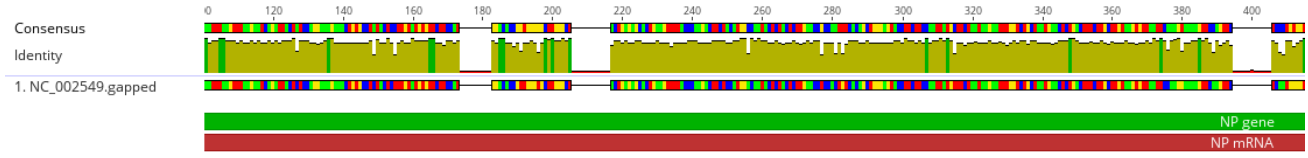


Figure 6. Identity scores visualization for each nucleotide position in a partial view of *Ebolavirus*' NP gene. Green zones have 100% identity, green–brown zones have between 30–100% identity, and red regions have less than 30% identity. Retrieved from Geneious version 2025.1 created by Biomatters. Available from https://www.geneious.com.(Geneious Prime, *2025b)*.

If more than 30% of the column is made up of gaps, the function considers the PPI for that position to be 0. This 30% threshold was chosen to ensure that the calculated pairwise identity (PPI) meaningfully reflects conservation among the actual sequence data, rather than being skewed by positions with insufficient information. Columns with excessive gaps are less reliable indicators of true sequence similarity, as the high proportion of missing data could artificially inflate or obscure identity scores. By discarding such columns, the function maintains the accuracy and biological relevance of the conservation analysis. Otherwise, the function calculates the amount of identical base pairs in proportion to the total number of possible pairs. Exact matches (e.g.: "A" versus "A"), are scored fully, whereas pairs

with ambiguous characters (e.g.: "A" versus "R") are scored based on the ambiguity matrix described above.

To determine the PPI score for a specific primer, its start and end positions were used to retrieve the corresponding alignment window, where each column was evaluated by the "pairwise_identity()" function, returning the average of all identity scores, divided by the number of valid columns (columns not completely comprised of gaps). This value was then converted into a percentage and appended to the primer information as its PPI.

Because the conservation of the nucleotides at the 3' end of a primer is extremely crucial to the binding to the target (Rychlik, 1995), the PPI score for the last three bases of each primer was calculated as its own parameter – PPI3'. While the calculation logic was the same as for the PPI function, the coordinates for the last three nucleotides depended on the type of primer. Forward primers had their coordinates extracted directly, with the index for the starting position being the end position minus three bases. However, for reverse primers, the binding orientation is opposite to the reference strand, meaning that the 3' end nucleotides are in the leftmost region of the alignment window of the sequence, requiring that this region be reverse complemented. This ensures that the bases assessed for reverse primers correspond to the functional 3' terminus.

Given that previous versions of this function used a different logic, it was necessary to validate the outputs, as they differed from the original function's results when tested on the same batch of primers. To do so, a random batch of 10 primers was chosen, and, with their coordinates, a cross-check of the obtained PPI values was performed with Geneious. By selecting the same alignment window in Geneious, the statistics tab automatically outputs a PPI score, which served as reference for evaluating both methodologies.

Once all primers were assigned their validated values for Percentage of Pairwise Identity (PPI) and Percentage of Pairwise Identity at the 3' end (PPI3'), the Conservation Score for each primer was computed. This involved taking the arithmetic mean of the PPI and PPI3' values for each primer, following the method defined by Carneiro et al., 2023. The resulting Conservation Score was then added to the primer information dictionary, ensuring that each primer had a comprehensive record of its sequence conservation metrics for subsequent analysis and selection steps.

2.4.3. Generating primer pairs

To only generate primer combinations that already meet some "best practices" standards for primer pair design, simultaneously cutting down computational resources to allow for the script to run on most machines, the following filters were applied prior to generating the combinations:

- Conservation Score threshold;
- Minimum and maximum values for GC content %;
- Minimum and maximum values for melting temperature.

The conservation score threshold was picked based on the size of the primer dataset, with 95% being the default value for literature primers (less quantity, easily computable) and 99% being the minimum for *in silico* ones (usually in the hundreds of thousands, prone to running into memory errors with more permissible values). GC content % and Tm intervals were established by cross-referencing literature values with histograms plotted from the datasets, showing the distribution of the primers for each parameter.

After applying the filters, primers were divided into two sets, Forward and Reverse. Given that a primer pair is inherently ordered (the forward primer is upstream of the target region, the reverse primer is downstream), the possible combinations can be defined by a Cartesian product:

$$F \times R = \{(f, r) | f \in F \; and \; r \in R\} \quad (4)$$

Where F is the set of forward primers, R is the set of reverse primers, f is an element of F and r is an element of R. To apply this logic to a data frame, each set had a column added, "key", with a set value of 1. To generate all possible primer pairs for further analysis, the Forward and Reverse datasets were merged using this common key, present in every row of both datasets. This merging process ensures a comprehensive pairing, whereby each forward primer is systematically matched with every reverse primer in the dataset. As a result, the merged data frame represents the complete set of all ordered pairs of forward and reverse primers. This exhaustive combination forms the basis for subsequent filtering and selection steps, allowing for the identification of primer pairs that meet the specified design criteria.

Finally, a second round of filtering is then performed to ensure that the generated pairs follow additional guidelines:

- Verify that, within each primer pair, the forward primer is positioned upstream relative to the reverse primer in the target sequence;

- Filter by amplicon length – the chosen length interval of the resulting amplicon, considering the scope of the project, was set to be between 70 – 1000 bp, spanning the recommended lengths for both standard (200 – 1000 bp) (Dieffenbach et al., 1993) and quantitative/real-time (RT-PCR, 75 – 150 bp) PCR (Applied Biosystems, 2005; Bio-Rad Laboratories, 2006; Debode et al., 2017);
- The difference between the melting temperatures of the primers should not exceed 5°C (Behind The Bench Staff, 2019; Geneious Prime, n.d.-a);

2.4.4. Parameter calculations: hairpin, homodimer and heterodimer formation

To evaluate the formation of undesirable secondary structures, such as hairpins (hereby referred to as self-fold or self-score), homodimers (homo-fold or homo-score) and heterodimers (dimer-fold or dimer-fold-score), the Primer3 module was used. Each primer was given a self-fold and a homo-fold value, whereas dimer-fold is a parameter exclusive to the corresponding primer pair, so it was only attributed to the combinations dictionary.

To prepare the pairs for evaluation of their No-Fold% scores, the final values for the self-score, homo-score, GC content % and Tm parameters were set as the average of each primer's corresponding values.

2.4.5. No-Fold%: scoring algorithm

Since Primer3 outputs scores as Gibbs free energy difference (ΔG, kcal·mol^{-1}) values, ΔG intervals were established as the "optimal conditions", for their respective parameters:

- Self-fold (hairpin formation) values should range from -2 to 0 kcal·mol^{-1}, to minimize both internal and 3' end hairpins (F. Carneiro et al., 2023; Pinheiro et al., 2024; Premier Biosoft, 2025);
- Homo-fold (homodimer formation) and Dimer-fold (heterodimer formation) values should range from -5 to 0 kcal·mol^{-1}, to minimize both internal and 3' end self and cross dimers (Benchling, n.d.; F. Carneiro et al., 2023; Pinheiro et al., 2024; Premier Biosoft, 2025; Sigma Aldrich, n.d.).

In this context, we considered the ΔG values to mean the following:

- ΔG is strongly negative: indicates that the primer-primer interactions are thermodynamically favorable, making dimerization or folding more likely;

- ΔG is closer to zero: indicates that the reaction is thermodynamically unfavorable, meaning undesirable secondary structure formation is more unlikely, and favors primer-target binding;
- ΔG = 0 corresponds to the equilibrium state between bound and unbound states.

In previous versions of the Conservation Score scripts, the transformation of the ΔG values into a 0–100 percentage range was done using the functions:

$$Score(\%) = \begin{cases} 0.05x + 100, & -2000 \leq x \leq 0\ cal \cdot mol^{-1} \\ 0, & otherwise \end{cases} \quad (5)$$

For self-fold% (hairpin formation) calculations, and:

$$Score(\%) = \begin{cases} 0.02x + 100, & -5000 \leq x \leq 0\ cal \cdot mol^{-1} \\ 0, & otherwise \end{cases} \quad (6)$$

For homo-fold% and dimer-fold% calculations.

This is a piecewise linear function with hard cutoffs, that can be improved to better represent a biological system. Molecules do not tend to behave in hard cutoffs, and primer parameters, such as binding affinity and secondary structure stability, are more prone to degrading gradually, not instantly, when hitting a threshold. Furthermore, anything outside of the "optimal conditions" interval is penalized equally, which does not allow for relative ranking of "bad" primers. For example, a primer that forms a hairpin with a ΔG of -2001 cal·mol^{-1} is scored the same as one with a ΔG of -5000 cal·mol^{-1}, when the first one is barely outside of the "ideal" range whereas the second might be catastrophically unusable.

Given this, a new, non-linear scoring algorithm was proposed, that aimed to have both a better statistical and thermodynamics foundation behind it.

The relationship between ΔG and the equilibrium constant K is given by the following equation:

$$\Delta G^{\circ} = -RT \cdot \ln(K) \quad (7)$$

Where R is the universal gas constant (≈ 0.001987 kcal·mol^{-1}·K^{-1}) and T is the reaction temperature in Kelvin, approximated by using the annealing temperature, $T_{annealing} \approx T_m + 273.15$. Although each individual primer has its own melting temperature, the restriction for the temperature differences

between primers in a pair ($\Delta T_m \leq 5°C$), implemented in the Primer Combinations step, ensures that the real annealing temperature does not deviate significantly from either oligonucleotide, given that experiments usually set the T_a of the reaction 2-5°C above the T_m (Ebertz, 2022). Rearranging the formula:

$$K = e^{-\Delta G°/RT} \quad (8)$$

Where K represents the ratio of bound to unbound primers, meaning that a more negative ΔG corresponds to a higher equilibrium constant, reflecting an increased probability of hairpin, homodimer and/or heterodimer formation.

If we consider the reaction $Primer\ (P) \leftrightarrow Dimer\ (D)$, the probability of a primer being bound in a dimeric state (P_{bound}) can be written as:

$$P_{bound} = \frac{[D]}{[P] + [D]} = \frac{K}{1 + K} \quad (9)$$

Substituting by the previous definition of K:

$$P_{bound} = \frac{e^{-\Delta G°/RT}}{1 + e^{-\Delta G°/RT}} = \frac{1}{1 + e^{\Delta G°/RT}} \quad (10)$$

This yields a sigmoid-like ($S(x) = \frac{1}{1+e^{-x}}$) relationship between ΔG and the probability of binding occurring, where:

- At ΔG = 0, $P_{bound} = 0.5$ (equal probability of bound and unbound);
- At ΔG strongly negative, $P_{bound} \to 1$ (almost all primers bound);
- At ΔG strongly positive, $P_{bound} \to 0$ (almost all primers unbound).

To express the primer's quality as its probability of remaining unbound (therefore free to bind to the target), we defined a score:

$$Score(\%) = (1 - P_{bound}) \times 100 \quad (11)$$

Thus, scores close to 100% translate to primers with reduced risk of dimerization/hairpin formation, whereas scores near 0% suggest that the primer is a poor candidate for experiments.

Although the logistic relationship between ΔG and primer dimerization probability arises naturally from thermodynamics, the sharpness of this transition in real PCR systems is modulated by additional factors (e.g.: ion concentration and primer concentration). So, to accommodate this, a threshold ($\Delta G_{threshold}$) was introduced to define the point at which primers should start being heavily penalized, as well as a slope (s) parameter that allows for the control of the penalty function's steepness. This second parameter allows the score to better reflect the gradual degradation of primer quality observed experimentally. Importantly, the slope can be calibrated to experimental data, ensuring the model remains both thermodynamically grounded and empirically robust. Adding these parameters to our previous expression to define the Penalty function:

$$Penalty = \frac{1}{1 + e^{\left(\Delta G - \frac{\Delta G_{threshold}}{2}\right)/(s \cdot RT)}} \tag{12}$$

And applying it to the Score function yields:

$$Score = (1 - Penalty) \times 100 = \frac{1}{1 + e^{-\left(\Delta G - \frac{\Delta G_{threshold}}{2}\right)/(s \cdot RT)}} \times 100 \tag{13}$$

Which resembles a logistic model:

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} \tag{14}$$

Where L = 1 is the curve's maximum, x_0 is the x value of the sigmoid midpoint ($\frac{\Delta G_{threshold}}{2}$, since threshold is approximating 0), and k is the logistic growth rate (steepness) of the curve ($-\frac{1}{s \cdot RT}$).

This model, unlike the previous one, implements softer cut-offs, allowing for a gradual decay in the score, while keeping the option to score primers more drastically with the introduction of the slope parameter. Another advantage is the ability to better discriminate primers by not lumping in primers slightly outside the threshold values with ones that are unequivocally unusable. This is especially advantageous when working with datasets with many "sub-optimal" candidates, due to its capacity to rank primers in a more relative and nuanced manner. Additionally, the new model resembles actual biological systems more

accurately, namely by introducing more "real" variables, such as the approximate temperature of the system at the time of the reaction. This allows, for example, two primers with similar ΔG values to be scored differently depending on the annealing temperature of the reaction, which is ignored by the linear mapping version. In sum, the new algorithm is built upon the fundamentals of thermodynamics while also providing a more robust probabilistic interpretation of the score, with no significant difference in processing times.

After applying this logic to obtain the scores (%) for each of the three parameters – self-fold score, homo-fold score, dimer-fold score – the No-Fold% was calculated using the expression:

$$No\text{-}Fold\% = \frac{\frac{Self_{forward}\% + Self_{reverse}\%}{2} + \frac{Homo_{forward}\% + Homo_{reverse}\%}{2} + Dimer\%}{2} \quad (15)$$

2.5. ViruScope application

2.5.1. Source code

All the scripts used to handle data throughout the course of the project were compiled, annotated and published on GitHub under the name "ViruScope". The repository contains four main scripts: viruscopeCLI.py (the full pipeline in a single script), and the three main components of the pipeline in individual files – arolit.py (includes all functions for the AROLit workflow), isopcs.py (all functions for iSOP, Primer Combinations and Conservation Scores) and utilities.py (auxiliary functions for quick data transformation in case the user needs to reformat inputs for the main functions). An extra script, timeoutwrapper.py, is also included and its function is to wrap the PDF converter function ("extract_from_pdf()") and make it, so the process is not infinitely stuck on a single file, skipping it after 5 minutes if extraction is not completed by then.

2.5.2. GUI implementation with Shiny

To make the tool more intuitive and easier-to-use for people with limited command-line experience, the source code was slightly adapted so it could be implemented in Shiny for Python, a module that allows for the creation of an application with a GUI. The implementation of the ViruScope graphical user interface using Shiny necessitated several adaptations to the project's core functions. Specifically, changes were made to how certain functions accepted inputs and delivered outputs to ensure compatibility with the interactive, session-based nature of Shiny.

To address potential crashes caused by memory allocation issues, a custom script was developed for creating temporary directories. These directories are used to store data while a session is active, with data being written directly to disk and retrieved as needed. Upon the conclusion of a session, the script performs a clean-up of the temporary directories, effectively preventing an accumulation of files within the main project directory.

The design phase incorporated various tools and technologies to enhance both functionality and aesthetics. Adobe Illustrator was utilized for tasks such as logo design and editing scalable vector graphics (SVGs). For the creation of custom, responsive user interface elements, a combination of HTML, CSS, and JavaScript scripts, was employed.

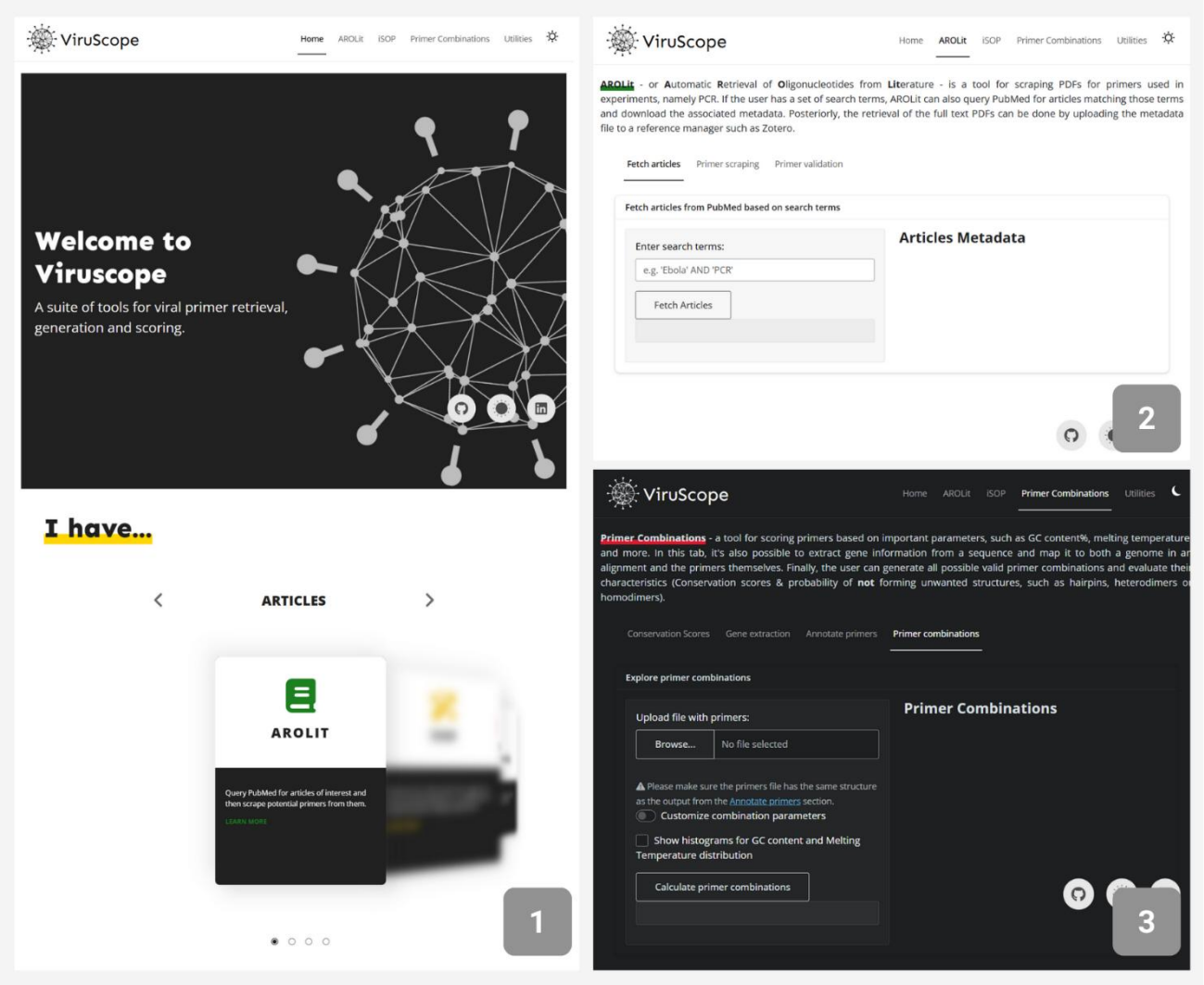


Figure 7. Screenshots taken of some of the ViruScope application's tabs. Image 1 is a full screenshot of the app's landing page, the Home tab, where the user can find the full logo, the navigation tab, the dark mode toggle, a small description of the tool, and a carousel with a description of each tool, along with a suggestion of which to choose based on the type of data ready for analysis, to guide first time users. Image 2 is an example of one of the tool tabs, with a concise example of what the tasks in that tab can be used for. Image 3 showcases a different tab with inputs of diverse types and serves as an example of the dark mode toggle.

The GUI version of ViruScope was also made available on GitHub, as ViruScope-GUI, along with installation instructions, requirements and a folder with example outputs for demonstration.

2.6. ViruScopeDB

The final datasets were exported to CSV and formatted into Excel tables to be ready for implementation. The website is hosted by Amen.pt on WordPress, and it consists of a landing page, an "About" section that contextualizes its purpose, a "Browse Viruses" tab that lists each virus and links to their AROLit and iSOP result pages, and a "Resources" page, where the user can access the ViruScope tool repository, read documentation, and learn how to cite both tools.

3. Results and discussion

This section aims to show the outputs of each step of the workflow, as the final product of this project is the ViruScope tool and its database, ViruScopeDB. To avoid redundancy, given the fact that the workflow was designed to be the same across all viruses, the results pertaining to analysis that were run individually for each virus – like AROLit, iSOP, combinations and scoring – are going to use the Ebola dataset outputs for exposition purposes. The remaining data associated with the other two viruses not present here can then be visualized directly on ViruScopeDB, if so desired.

3.1. Genomic data collection and analysis

3.1.1. Retrieval of sequences

The summary of the results for the NCBI Virus database search can be found in **Table 2**. A post collection step was performed to remove sequences that were completely identical, minimizing the redundancy of subsequent analyses.

Table 2. Summary of the filtering parameters applied to each virus during the NCBI Virus database search, including criteria such as nucleotide completeness, sequence length, and allowable ambiguous characters, as well as the total number of sequences extracted and the final count after removing duplicates.

Virus group	Virus/ TaxID	Nucleotide completeness	Minimum length	Maximum length	Max ambiguous characters	Sequences extracted	Number of sequences after removing duplicates
HIV	Human immunodeficiency virus 1, taxid:11676	Yes	-	-	-	7,261	6,613
	Human immunodeficiency virus 2, taxid:11709	Yes	-	-	-	43	39
Ebola	Bundibugyo virus, taxid:56995	Yes	-	-	-	658	489
	Sudan ebolavirus, taxid:186540	Yes	-	-	-		
	Taï Forest ebolavirus, taxid:186541	Yes	-	-	-		
	Zaire ebolavirus, taxid:186538	Yes	-	-	-		

	Ebolavirus, taxid:1570291	Yes	-	-	-		
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2, taxid:2697049	Yes	29800	30000	0	321916	

3.1.2. Sequence alignments and general statistics

The general statistics for each alignment file were collected directly from Geneious Prime's "Statistics" tab and are detailed in **Table 3**. Taking Ebola as an example, the mean sequence length across all sequences was 18,921 bp, slightly below yet still very similar to the reference sequence, NC_002549.1, own size of 18,951 bp, suggesting that the genome is very uniform in length even across species. The average GC Content % of the alignment (41.3%) was also consistent with the reference sequence's, 41.1%, reflecting highly conserved base composition. Since PPI measures the average nucleotide identity between any two sequences, a value of 91.5% for this parameter reveals that the different sequences diverge in 8.5% of their nucleotide compositions, which can be attributed to intraspecies (e.g.: Zaire vs. Sudan ebolaviruses) variations. Thus, we can affirm that the dataset captures the genetic diversity across the *Ebolavirus* genus. PIS, or the Proportion of Identical Sites, represents the percentage of alignment positions that are identical across all sequences analyzed. In this context, the PIS value indicates that only around 30% of the positions are fully conserved, meaning that approximately 70% of positions display at least one sequence that differs from the others in the alignment. This relatively low PIS value can be attributed to the inherent diversity present when comparing a large number of sequences—489 in total.

Despite this observed divergence, it is important to note that about one quarter of the entire genome remains conserved regardless of the species or strain. This conserved segment forms the genomic

backbone of the genus. Therefore, it is expected and natural to observe some degree of sequence variation when such a comprehensive comparison is conducted.

Table 3. Summary of key alignment metrics generated by Geneious Prime for each viral dataset, including mean sequence length, GC content, proportion of identical sites (PIS), and pairwise percent identity (PPI).

Virus	Mean sequence length (bp)	GC Content (%)	PIS (%)	PPI (%)
Ebola	18,921	41.3	28.3	91.5
HIV-1	8,891	41.6	0.04	80.6
HIV-2	10,088	45.6	38.1	80.8
SARS-CoV-2	29,835	37.9	73.1†	99.8

†Statistic obtained by random sampling of 3000 sequences from an alignment file of 285,921 sequences. Sampling was necessary due to repeated crashes while trying to calculate the value. The sample's other statistics corresponded to the full set values, indicating it can be interpreted as representative of the full alignment.

3.2. Gene extraction and mapping

Gene extraction and mapping are crucial steps in analyzing viral genomes, allowing researchers to identify and verify the locations of coding sequences within a reference. In the context of this study, the gene extraction function was applied to Ebola's reference genome to demonstrate its effectiveness and accuracy. Using Ebola as an example of use and validation for the gene extraction function, it was possible to confirm that all genes that make up the reference genome were extracted. The coordinates in the reference sequence also matched the positions detailed in its NCBI entry. Additionally, alignment coordinates were also successfully mapped, with all positions matching the ones calculated using the Geneious Prime's premium feature for the same purpose, "Annotate from Database" (**Table 4**).

Table 4. Example of an output of the gene extraction and mapping functions. Using Ebola's RefSeq ID as the input, NC_002549.1, the function extracted all the CDS from the retrieved GenBank file and returned the positions of each gene in the reference and the alignment.

Gene	Reference start	Reference end	Alignment start	Alignment end
NP	55	3026	56	4574
VP35	3031	4407	4580	6049
VP40	4389	5894	6032	9692
GP	5899	8305	9698	12684
VP30	8287	9740	12667	14314
VP24	9884	11518	14508	16394
L	11500	18282	16377	23365

3.3. AROLit database

3.3.1. Retrieved articles

From the queries submitted through Entrez to the PubMed database, a total of 263, 6,169 and 16,670 PCR-related articles for Ebola, HIV and SARS-CoV-2, respectively, were returned in Medline format. After uploading the data files to Zotero and requesting the automatic retrieval of the full PDFs, Zotero managed to find and store 98% of the documents for Ebola, 42% for HIV, and 59% for SARS-CoV-2 (Table 5).

Table 5. Number of articles from which metadata was retrieved from PubMed using Entrez versus the amount of full text PDFs automatically found and downloaded by Zotero.

Virus	Articles found (Entrez)	PDFs found (Zotero)
Ebola	263	258
HIV	6,169	2,579
SARS-CoV-2	16,670	9,886

The drop in the retrieved metadata to retrieved full text ratio is directly linked to the scale of the submitted dataset. The sharp increase in the number of articles makes the retrieval of documents a lengthier and more resource intensive process, sometimes triggering CAPTCHA tests due to the high volume of requests being made to the databases or even causing Zotero to crash. The amount of articles effectively found for HIV and SARS-CoV-2, however, is still big enough for a significant amount of information – several times more than the amount included in HIVoligoDB and COV2ID – to be collected and processed.

3.3.2. Evaluation of primer scraping methods

To ensure a comprehensive comparison of strategies for extracting primer sequences from scientific literature, several methods were systematically evaluated in this study. As described in the methodology section, the main three methods tested consisted of two LLM-based approaches, one run locally (DonutAI/OpenLLaMa-7b, through the HuggingFace module for Python) and the other run on the developer's website (Google's NotebookLM). Running the models locally revealed itself to be too computationally expensive for the hardware available, in terms of CPU requirements, even for the downscaled versions (less parameters), and training and/or tweaking the models into a fast, highly accurate and deployable version would be too time-consuming for the scope of the project. NotebookLM, while useful for smaller datasets and more direct, context-requiring inquiries, turned out to scale badly as the number of PDFs in the test batch grew, and it was unable to output reproducible results. The results from a run performed on a "fresh" model (memory of previous chats deleted to avoid bias and "cross-contamination" between attempts) only managed to correctly identify 47 out of 119 primers, and repeat prompting further worsened the results. A more traditional approach using PyMuPDF4LLM and regular expressions scored the best out of all the tested approaches, correctly identifying 83 out of 119 primers (**Table 6**).

Table 6. Comparison of the performance and scalability of three primary primer extraction methods evaluated in this study.

Model	Accuracy (%)	Direct pipeline integration?	Execution time (s)
RegEx + PyMuPDF4LLM	69.7	Yes	41.8
DonutAI/OpenLLaMa-7b	0	Yes	300+
NotebookLM	38.84	No	67.2

An analysis of the primers that evaded extraction revealed that the problem lay with the conversion step in the vast majority of cases. More complicated formatting (e.g.: tables in image format, multi-column PDFs, low quality documents, less legible fonts) often resulted in the primers not being extracted at all or

being cleaved in ways that made it so only parts of the sequence were captured. For the validation step, correct identifications were only attributed to cases where the primer was extracted in its entirety. In practical applications, primers may still function effectively even when one or two mismatches are present. The impact of these mismatches largely depends on their location within the primer sequence. Notably, mismatches occurring at the 3' end are more likely to significantly affect the reaction outcome (Koehler et al., 2023; Stadhouders et al., 2010). This observation helps explain instances where primers with only a single nucleotide discrepancy may not perform as expected, despite otherwise matching their intended targets.

The inclusion of a BLAST validation step further underscores the robustness of the AROLit methodology. By systematically assessing primer–target alignments, this approach ensures that only sequences with sufficient specificity are retained, thereby accounting for and minimizing incorrect captures due to minor mismatches.

3.3.3. Case study: AROLit applied to Ebola articles

3.3.3.1 General statistics

From the 258 articles obtained from Zotero, the AROLit scraping algorithm managed to extract 921 sequences, of which 174 were validated as being Ebola primers. After applying the sequence expansion function for primers with International Union of Pure and Applied Chemistry (IUPAC) ambiguous nucleotides, the final number of unique primers in the dataset was 476. Of the retrieved sequences, 53% were forward primers and 47% were reverse primers (**Figure 8**).

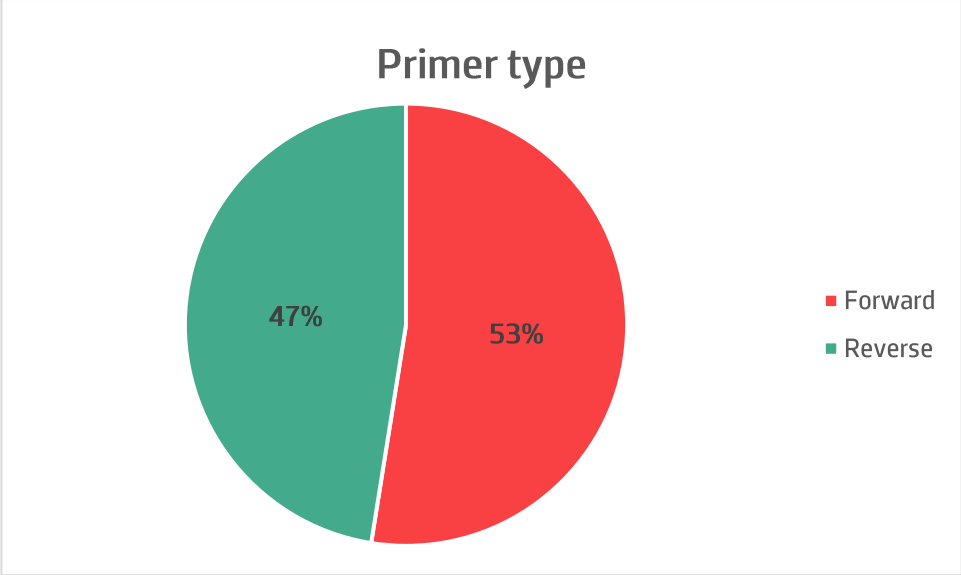


Figure 8. Distribution of primer orientation within the AROLit dataset: number and percentage of forward primers (250; 52.5%) and reverse primers (226; 47.5%) among a total of 476 unique validated sequences.

In terms of which regions the primers bind to, the most targeted gene is the L (Polymerase) gene (29%), closely followed by the Nucleoprotein (NP) and Glycoprotein (GP) encoding genes, both at 25% (**Figure 9**). The 14% attributed to intergenic/unknown regions are mostly attributed to the border region between VP40 and the GP genes, encompassing VP40's polyadenylation site, responsible for signaling the end of transcription for the VP40 gene, a short intergenic region, and the GP transcription start signal (Sanchez et al., 1993).

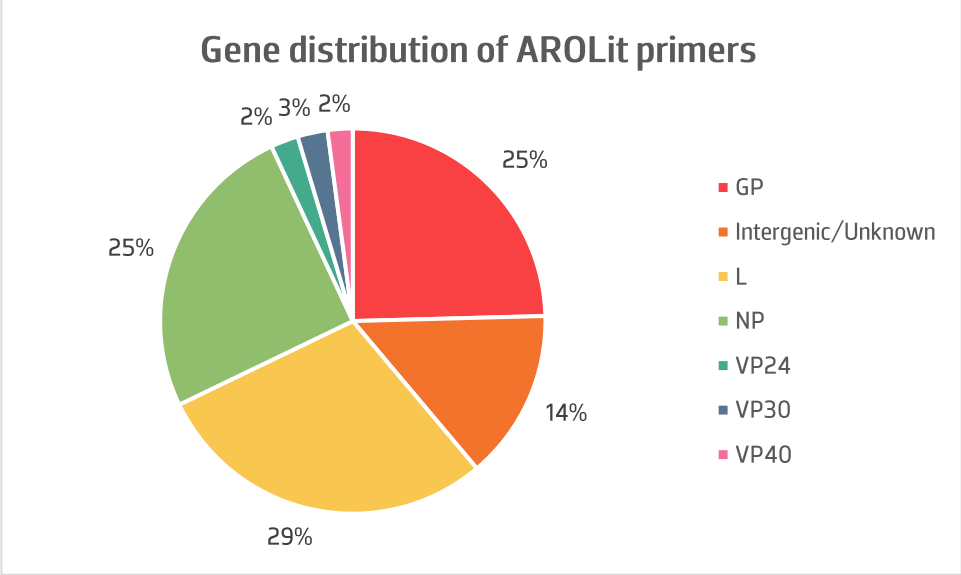


Figure 9. Distribution of Validated Ebola Virus Primers by Genomic Region. Proportion of unique, validated primers extracted using the AROLit methodology, categorized by their binding sites within the Ebola virus genome. The figure details the percentage of primers targeting the L (Polymerase), NP (Nucleoprotein), GP (Glycoprotein), VPs (Viral Proteins) and intergenic/unknown regions, reflecting the preferential selection of conserved and diagnostically relevant loci.

These results are consistent with both the available information in the literature for the most targeted regions in Ebolavirus detection methods, and the high degree of conservation observed in the sequence alignment file for these positions (**Figure 10**). RT-PCRs targeting the NP gene are the most widely used for detection and diagnosis, not only for Ebola and its many variants, but also for other filoviruses, due to their high specificity and sensitivity (Bell et al., 2015; Jääskeläinen et al., 2020; Trombley et al., 2010). Bell et al. (2015) also concluded that single nucleotide polymorphisms (SNPs) are not as commonly present in the NP regions most targeted by primers, in contrast to other genes such as GP, making the reduced variation in this zone another advantage for primer design. Despite this higher variability in the GP gene, it still constitutes a valuable target, with some studies obtaining higher sensitivities in assays targeting GP in comparison to those targeting NP (Yang et al., 2017). Similarly to the results obtained in the EbolaID project (J. Carneiro & Pereira, 2016b), the L gene primers make up the largest percentage of oligonucleotides designed for Ebola PCR experiments, and the three top scoring primers in the database were all L gene oligonucleotides and part of the best primer pairs. Furthermore, the conservation of this gene extends to other filoviruses, making it possible for an assay designed for this region to reliably detect all filoviruses known to cause disease in humans (Jääskeläinen et al., 2019). An additional point, originally noted in the EbolaID project and still evident nearly a decade later, is the persistent absence of literature oligonucleotides targeting the VP35 gene. This scarcity is particularly notable given that the VP35 region is relatively conserved, which would typically make it an appealing target for primer design. However, the lack of primers may be attributable to chance and historical research focus, early diagnostic efforts and published assays may have prioritized other genes (such as L, NP, and GP) that were perceived as more diagnostically relevant or for which validated protocols already existed. As a result, once these gene targets became standard, subsequent studies may have continued to build upon them, inadvertently overlooking the VP35 gene despite its conservation and potential suitability for reliable primer binding.

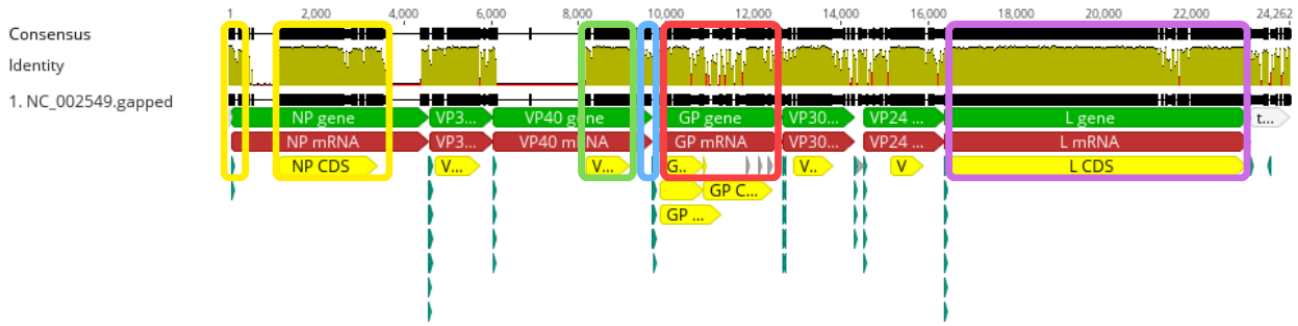


Figure 10. Identity scores for the regions identified as most targeted by literature primers. Every region shows a degree of conservation between 30–100%, as noted by the greeny-brown bars in the Identity row. Zooming in on these regions also reveals short sequences with 100% identity across all 489 genomes in the alignment, mostly in zones attributed to regulatory tasks. Highly conserved regions are marked for the NP (yellow boxes), VP40 (green), GP (red) and L (purple) genes, and the intergenic VP40–GP region is marked in blue. Retrieved from Geneious version 2025.1 created by Biomatters. Available from https://www.geneious.com

Exploring **Figure 10** further, we can see that the L gene is especially conserved, with several short motifs fully conserved across the entire CDS, explaining why primers may target it more frequently. The NP gene's conserved regions are not continuous, with the high identity sections representing the NP CDS and the first ~400 bp. A very highly conserved region that was not extracted to the Locus mapping was the leader 5'UTR (untranslated) region, located in the first 55 bp of the genome (leftmost yellow box). Highly cited articles for primer design best practices, like Dieffenbach et al.'s 1993 *General Concepts for PCR primer design*, advise against targeting the 5'- and 3'- untranslated regions due to their tendency to be less conserved across species. In the case of our alignment file, however, this 5'-UTR boasts an incredible 96.8% pairwise identity, with over half of the nucleotides (58%) being identical sites. VP40 and GP's regions are delimited by the green and red selection boxes, respectively. Most importantly, the intergenic region (blue selection box) shows a 100% identity score for the 11 bp VP40 regulatory and 12bp GP regulatory, marking these zones as fully conserved across all sequences.

3.3.3.2 GC% content

The general recommended GC content % interval for a primer is often quoted as being between 40–60% (Behind The Bench Staff, 2019; Bio-Rad Laboratories, 2006; Dieffenbach et al., 1993), though some sources recommend ranges as wide as 30–80%, with the optimal range being 45–55% (Applied Biosystems, 2005; Hall-Wheeler, 2015). Some sources even state that there is no interval at all, and GC content should instead depend on the GC content of the desired amplicon (Rychlik, 1995). Although these reference values are far from static, GC Content % that is too high is prone to forming secondary structures, while values that are too low affect stability and binding by lowering the melting temperature (Behind The Bench Staff, 2019; Hall-Wheeler, 2015).

To assess the performance of AROLit primer pairs within this framework, we analyzed the average GC content (%) for each primer type across different genomic regions, as illustrated in **Figure 11**.

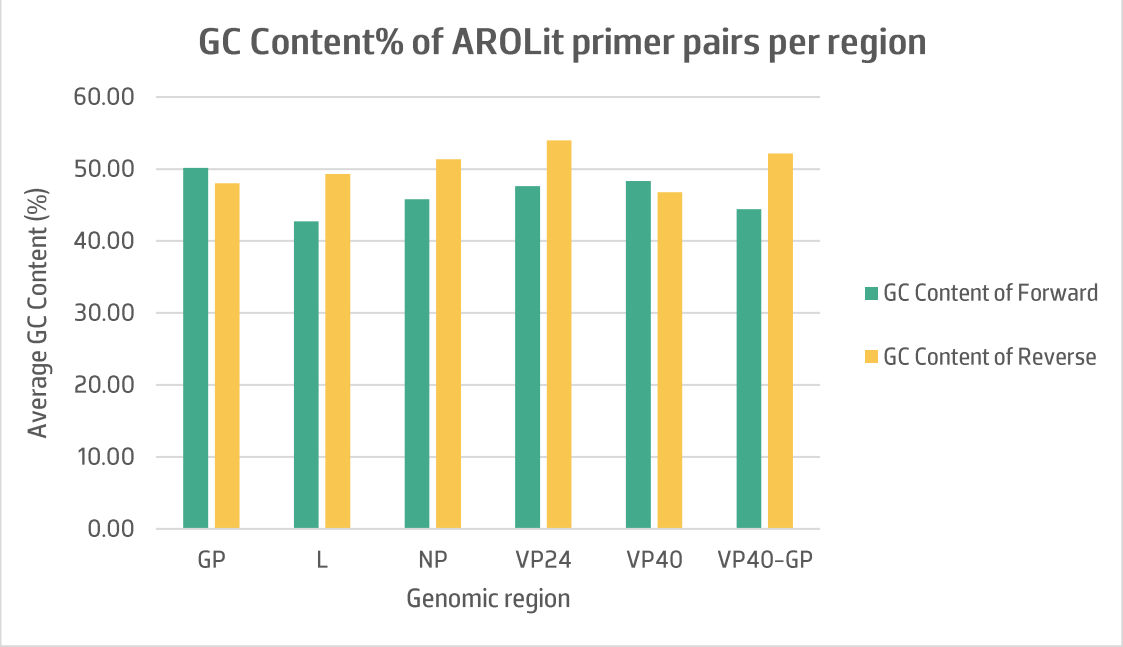


Figure 11. Average GC content (%) of forward and reverse primers in AROLit primer pairs, stratified by genomic region. Values represent the mean GC composition for each primer orientation within specific Ebola virus genome regions, as generated by the Combinations script. This analysis highlights primer stability and the influence of regional sequence composition on primer design parameters.

Across all genomic regions, the average GC Content % for all primers stays in the 40–55% range, which falls within the most commonly accepted interval for stable primers with reduced secondary structure formation. Although relatively well balanced, reverse primers tend to have slightly higher GC composition than forward primers, especially for those in the NP, VP24 and VP40–GP regions. This can be the result of composition bias within the regions themselves, with the consensus sequence's NP, VP24 and VP40–GP coding DNA sequences having GC Content percentages of approximately 46%, 43% and 42%, higher than the overall amount.

3.3.3.3 Melting Temperature

A similar analysis was conducted for the Tm values. Despite literature also not being concordant in regards to an ideal range, commonly accepted intervals include 58–60°C (Applied Biosystems, 2005), 50–65°C (Bio-Rad Laboratories, 2006), 65–75°C (Behind The Bench Staff, 2019) and 56–62°C (Dieffenbach et al., 1993). More importantly, primer pairs with a Tm difference higher than 5°C should be avoided (Behind The Bench Staff, 2019), with 1–2°C being the optimal difference if the same Tm cannot be guaranteed (Ebertz, 2022).

AROLit primer pairs fit nicely into these parameters, as can be observed in **Figure 12**.

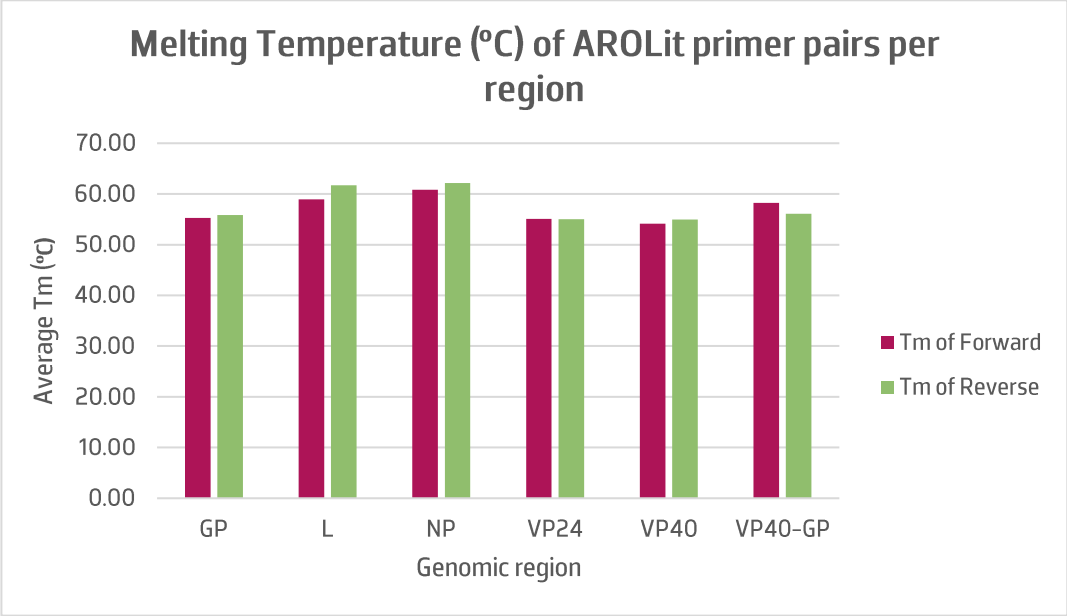


Figure 12. Average melting temperature (Tm, °C) for forward and reverse primers in AROLit primer pairs, stratified by genomic region. Data represent mean Tm values for each primer orientation, as generated by the Combinations script, highlighting regional differences and primer pair compatibility within the Ebola virus genome.

All primers have average Tm values in the 54–62°C range, which is nested within multiple cited "optimal" ranges. Additionally, most paired values are within 1–2°C of each other, which is the ideal difference between temperatures, with the exception of the L region, where the average difference is approximately 3°C. The Tm of the reverse primer is also generally higher than the forward's in all pair averages except in the VP40–GP region, which is consistent with the reverse primers' tendency to have higher GC Content percentages in the previous section.

3.3.3.4 Conservation Scores

Due to the increase in the amount of data being processed in this project, optimizations of the PPI (and, subsequently, PPI3') function were needed to make it scalable, faster and, if possible, more accurate.

Table 7. Comparison of the values obtained for each test primer using the old version of the PPI script, the revamped version described in the methodology section, and the reference values obtained from Geneious Prime. The new PPI script, along with being significantly faster even with larger datasets, is also more accurate and sensitive.

Primer sequence	Old PPI script	New PPI script	Geneious Prime
GAGAAAAGGCTTGCCTTGAG	91.0	95.1	95.4
CATGTGCATCCCTTGGTGTA	100.0	97.9	98.0
CCAACAGCTTGGCAATCAGTAGG	100.0	97.0	97.2
ATGCCGGAAGAGGAGACAA	100.0	96.0	96.2
GCAGAGCAAGGACTGAT	100.0	97.4	97.5
GTTCGCATCAAACGGAAAAT	100.0	96.2	96.4
TCTGACATGGATTACCACAAGATC	83.9	94.9	94.5
GCCAACGATGCTGTGATTTC	100.0	94.4	94.7
GGAGACGAACTCCTCGTTCTG	100.0	95.2	95.4
TAGTTAYTCGCACACAA	81.3	99.4	99.4

The new script, along with improving the accuracy of the scores (**Table 7**) and introducing support for degenerate bases, also reduced the runtime from multiple days to a few hours.

Regarding the score value, a higher conservation score (CS) means that the primer binds with high sequence identity across all alignment sequences (PPI) and the last 3 nucleotides of the 3′ end terminus are particularly well conserved (PPI3′), with the second parameter being especially relevant to PCR efficiency (Dieffenbach et al., 1993; Rychlik, 1995). Overall, the CS scores for the AROLit primers point towards a high degree of conservation across all regions, with values between 92–96% (**Figure 13**).

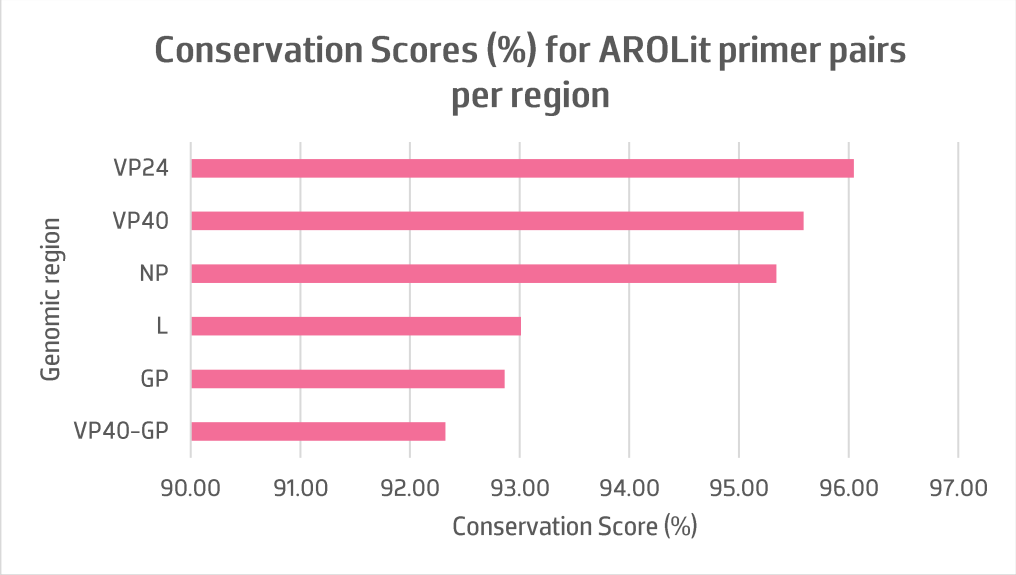


Figure 13. Conservation Score (%) averages for AROLit primer pairs, per genomic region.

Even though the L gene appears to be the most overall conserved gene – based on the consensus sequence's L gene's PPI score presenting as the highest among individual gene PPI – primer pairs targeting VP24, VP40 and NP have better CS scores. This apparent discrepancy between global and primer conservation scores can be explained by the fact that while the first case describes nucleotide variability across the entire gene, the primer scores only consider sequence identity at their specific binding sites. Thus, if the primers target less globally conserved genes but they all concentrate on small stretches of particularly high uniformity, this can lead to stronger scores. Conversely, given the size of the L gene and its status as a primary target, more primer pairs exist, and more sections of the gene are targeted. If local conservation at the binding sites of a few pairs is lower, that can skew the CS score average for that region. In fact, if we sort the primer pairs by their CS, highest to lowest, the first 8 results are all pairs targeting the L gene. On the other hand, evidence stating that VP40 and VP24 are potentially more conserved than other genes in the *Ebolavirus* genus can also be found in the literature (Carroll et al., 2015; Sanchez & Rollin, 2005).

3.3.3.5 No-Fold%

As part of refining primer design evaluations, it is crucial to assess not only conservation but also structural stability, an important factor in PCR efficiency. The No-Fold% score provides insight into a primer's propensity to avoid forming undesirable secondary structures, such as hairpins or dimers, which can negatively impact experimental outcomes. To illustrate how this metric responds to various thermodynamic parameters and algorithmic adjustments, we constructed four graphics that compare

the behavior of the No-Fold% score across different ΔG values and slope settings, as well as its performance relative to the previous scoring approach (**Figure 14**).

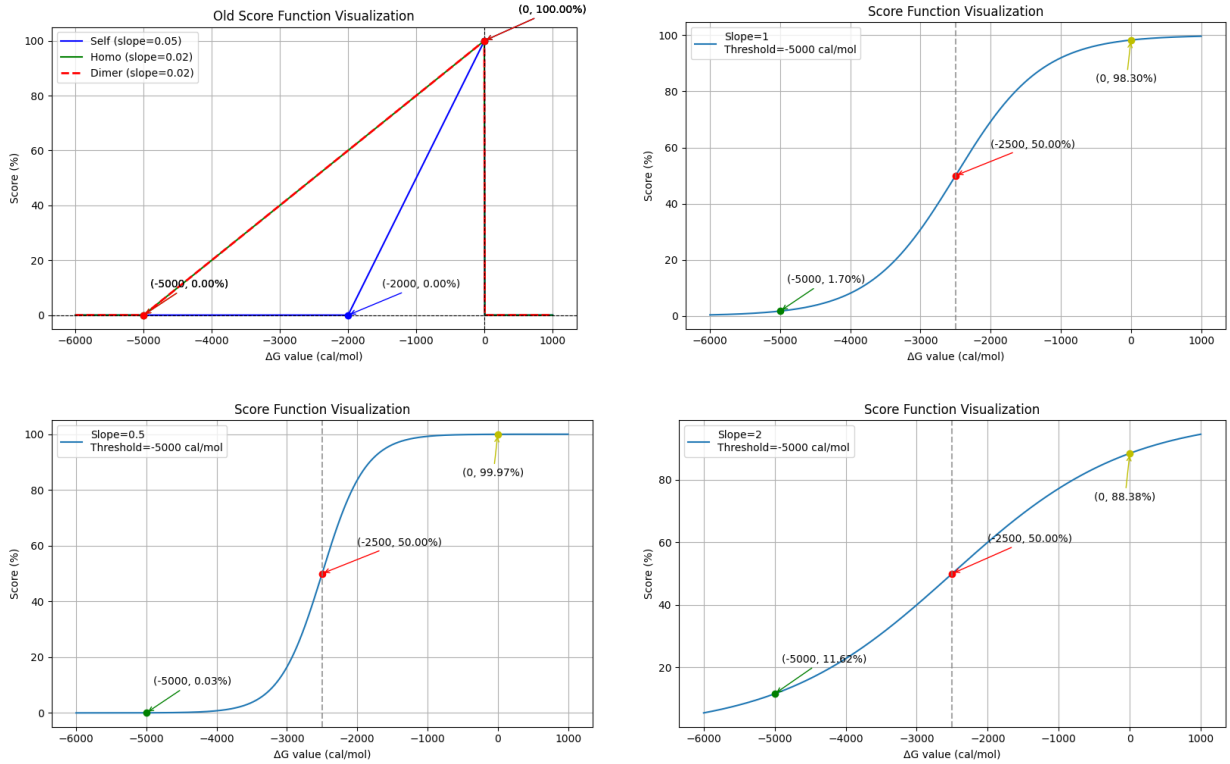


Figure 14. Graphical visualization of the old linear mapping score function (top left), the new logistic mapping function with the default slope value of 1 (top right), the new function with a slope value of 0.5 (bottom left) and the new function with a slope value of 2 (bottom right), All the graphics have labels for the x-axis values corresponding to the ΔG interval extremities (-5000 and 0, and one more for -2000 in the old function), and the corresponding y value given by the algorithm (the score). The graphics pertaining to the new algorithm also have the coordinates for the sigmoid mid-point, corresponding to half of the threshold value.

In comparison to the old scoring function, the new version fulfills the same principles (0-100% range, with ΔG values approaching zero scoring closer to 100%), while representing biological phenomena more realistically (non-linear function). The slope value can also be adjusted to be more or less permissible, depending on how "radically" a user wants to score a primer, especially when thinking in terms of the ViruScope tool, where the function can be used on private datasets that could benefit from further parameter customization.

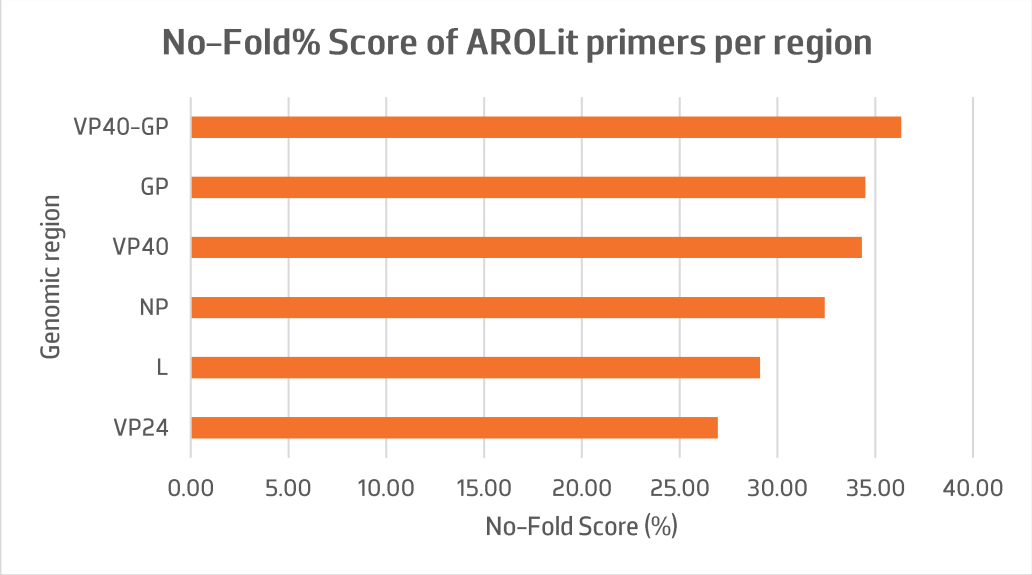


Figure 15. Average of No-Fold% Scores for AROLit primer pairs, per genomic region.

In the scope of this project, the scores were calculated using the default slope value (=1). The returned averages for No-Fold% scores varied between 27-36%, with the more stable primers belonging to the VP40-GP, VP40 and GP regions. Sorting primers from highest to lowest scores, we could determine that the most stable primer pair (61% score) targeted the GP gene, and originated from the same forward sequence, paired with some of the variants of a primer with multiple degenerated bases. Interestingly, only the variants with G and C bases substituting the ambiguous nucleotide at the end were present, which is not unexpected due to the importance of GC clamps at the 3' end of oligonucleotides (Applied Biosystems, 2005; Behind The Bench Staff, 2019; Benchling, n.d.; Bio-Rad Laboratories, 2006; Dieffenbach et al., 1993; Ebertz, 2022; Hall-Wheeler, 2015; Premier Biosoft, 2025; Rychlik, 1995; Sigma Aldrich, n.d.; Stadhouders et al., 2010). In general, this primer pair fulfilled all benchmarks for good primer design.

Although these values appear generally unfavorable at first, the No-Fold% score's individual parameters (Self-fold, Homo-fold and Dimer-fold) were already scored using a ΔG interval where all values within are considered well-tolerated for a PCR experiment (Sigma Aldrich, n.d.). Rather than meaning that the primer is unequivocally unusable, it means instead that it is on the lower end of the quality scale compared to those scoring higher. Taking the lower bound of the hairpin formation reference interval, ΔG = -3 kcal/mol, as a reference point, it simply means that more negative values may make the PCR reaction's temperature unable to produce enough energy to break the secondary structure and "free" the primer. However, -3 kcal/mol is still considered a well-tolerated value. No-Fold% scores are a comparative

measure between primers and primer pairs that are already viable, with more loose implications for primer quality in a global sense. This type of evaluation would require the creation of a model that accounted for more reaction variables, like primer concentration and polymerase characteristics.

3.3.4. Case study: iSOP tools applied to Ebola

3.3.4.1 Primer generation and validation for the alignment

Applying the generator function to the reference genome of *Zaire ebolavirus* (NC_002549), we obtained all possible sequences between 17 and 32 nucleotides in length, resulting in a total of 605,936 unique primers. Of these, only 572,872 produced hits in the BLAST against the alignment (**Table 8**). Given the sheer number of sequences left, the decision was made to not loosen the BLAST parameters further as we were left with a representative number of primers.

Table 8. Comparison between the total number of primers generated using "generate_primers()" and the number of primers left after running a BLAST against the genome in the alignment.

Number of generated primers	Number of primers after running BLAST	Eliminated primers
605,936	572,872	33,064

3.3.4.2 General Statistics

After generating and validating primers for the reference Ebolavirus genome using the iSOP pipeline, an analysis was conducted to determine the distribution of validated primers across different genomic regions. Similarly to what happened with AROLit, the region with most validated primers was the L gene, followed by the NP and GP genes. In contrast to the previous analysis though, the remaining genes seem to have a more equal distribution of primers, with VP35 showing up in the graph, the only gene that had no associated primers in the AROLit dataset (**Figure 16**).

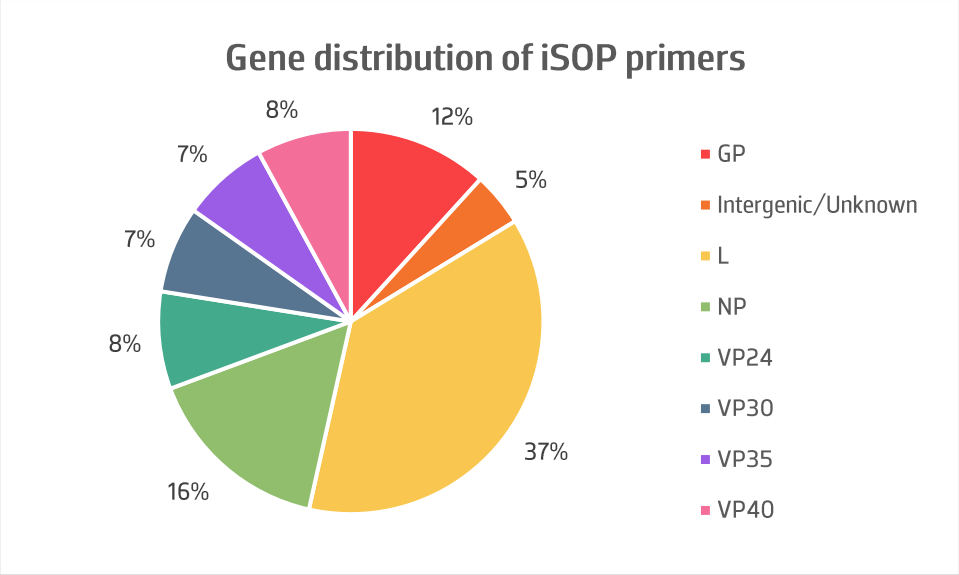


Figure 16. Percentage of validated primers generated by iSOP, per genome region.

The appearance of all genes and their distribution is expected, as primers for every region were generated systematically and not with specific targets in mind, and then filtered to keep those with good coverage across all sequences, creating a more balanced distribution. However, due to the large amount of data generated by iSOP, calculating every single possible primer pair would be computationally impossible (~82 billion combinations), therefore filters were applied prior to pairing up the sequences. As applying the default threshold of minimum 95% CS still made the combinations function run into memory allocation errors, further tightening of this parameter to only consider primers with CS ≥ 99% was applied. This filtering reduced the number of primers in the database to 2,916 and changed the gene distribution slightly (**Figure 17**).

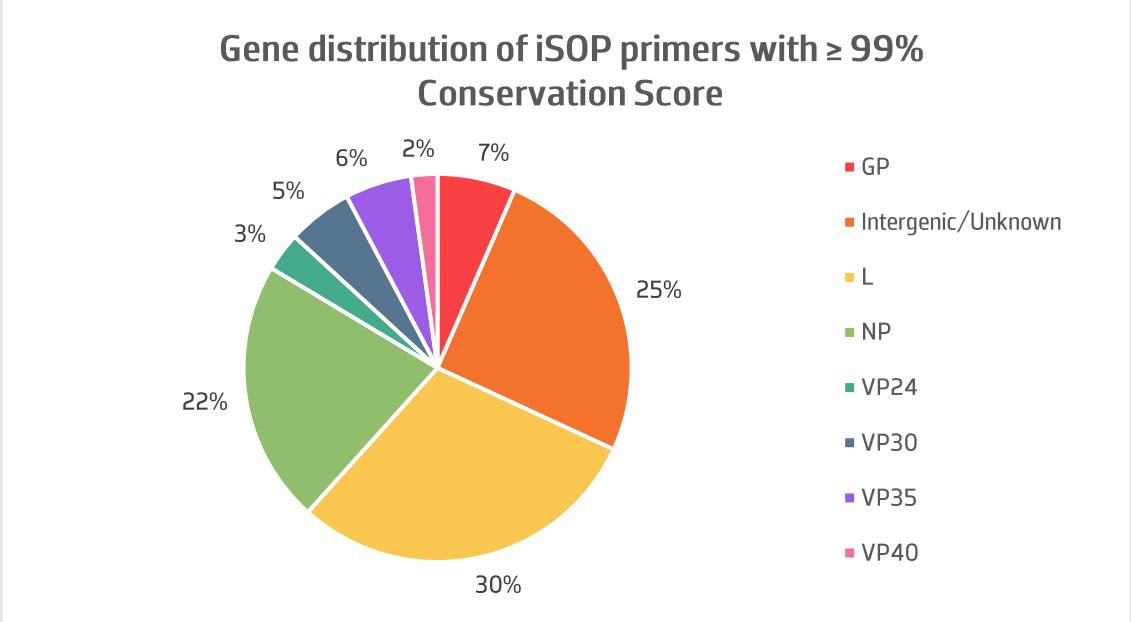


Figure 17. Percentage distribution of iSOP primers with Conservation Scores ≥99% across Ebolavirus genome regions, including L, NP, GP, VP35, and Intergenic/Unknown genes.

While every gene was still represented in the filtered dataset, key differences can be observed: a 20-percentage point increase of the Intergenic/Unknown portion of the dataset, revealing that these regions may be more highly conserved across species than expected; a drop in primers binding to the GP region, indicating that there are regions of this gene that are less conserved across species, which is consistent with literature findings (Carroll et al., 2015; Mahale & Patole, 2015; Sanchez & Rollin, 2005). Considering that viral glycoproteins are major targets for both immune response and therapeutics, the increased selective pressure exerted on GP can explain its increased variability.

3.3.4.3 GC% content

To confirm that the primers were within acceptable GC Content % ranges and visualize where most of them were concentrated, a histogram was made to determine the distribution of this parameter within the full dataset (**Figure 18**).

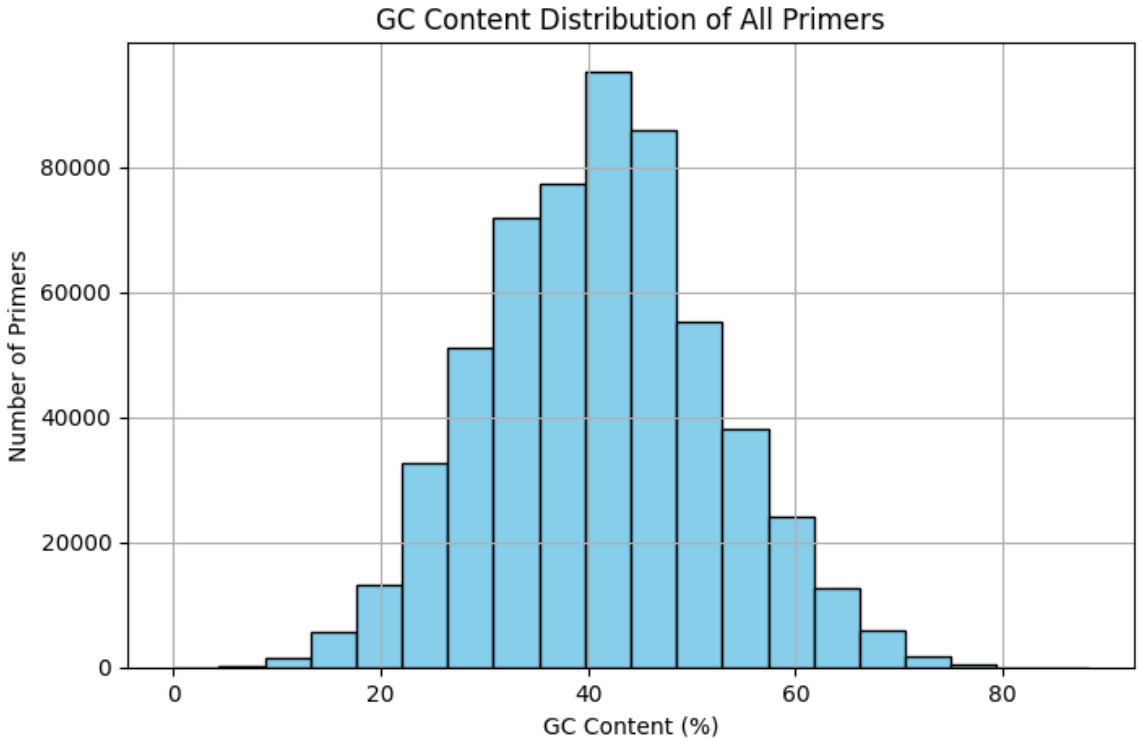


Figure 18. Histogram showing the GC content (%) distribution of all iSOP-generated primer sequences.

Most of the iSOP primer sequences concentrated around the 40-48% range, which encompasses the consensus sequence's GC Content % and falls within the accepted interval for primers, previously described in the AROLit section. It is also nested within the AROLit dataset's interval for this parameter. A posterior analysis of the GC composition of the generated primer pairs, which are exclusively comprised of sequences with a CS equal to or greater than 99%, also shows that throughout all genomic regions targeted, both primer averages for the parameter stay within the optimal 45-55%, and are relatively well balanced (**Figure 19**).

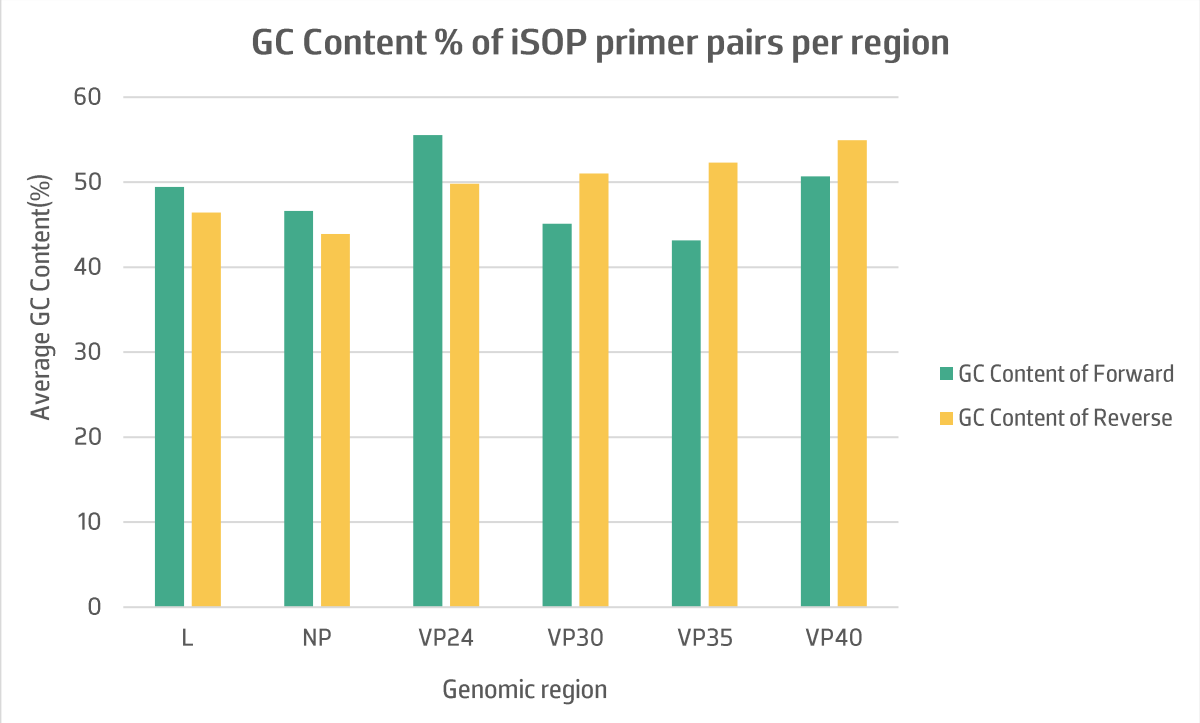


Figure 19. GC Content (%) average for each primer type (Forward and Reverse) in the iSOP primer pairs generated by the Combinations script, per genome region.

The final GC Content % averages for iSOP-generated primer pairs with 99%+ Conservation Scores, 45-55%, is remarkably similar to the interval obtained in the AROLit analysis (40-55%), indicating a good crossover between literature and computational simulations for this parameter.

3.3.4.4 Melting Temperature

A similar analysis was conducted for the Tm values, once again for the full dataset (**Figure 20**).

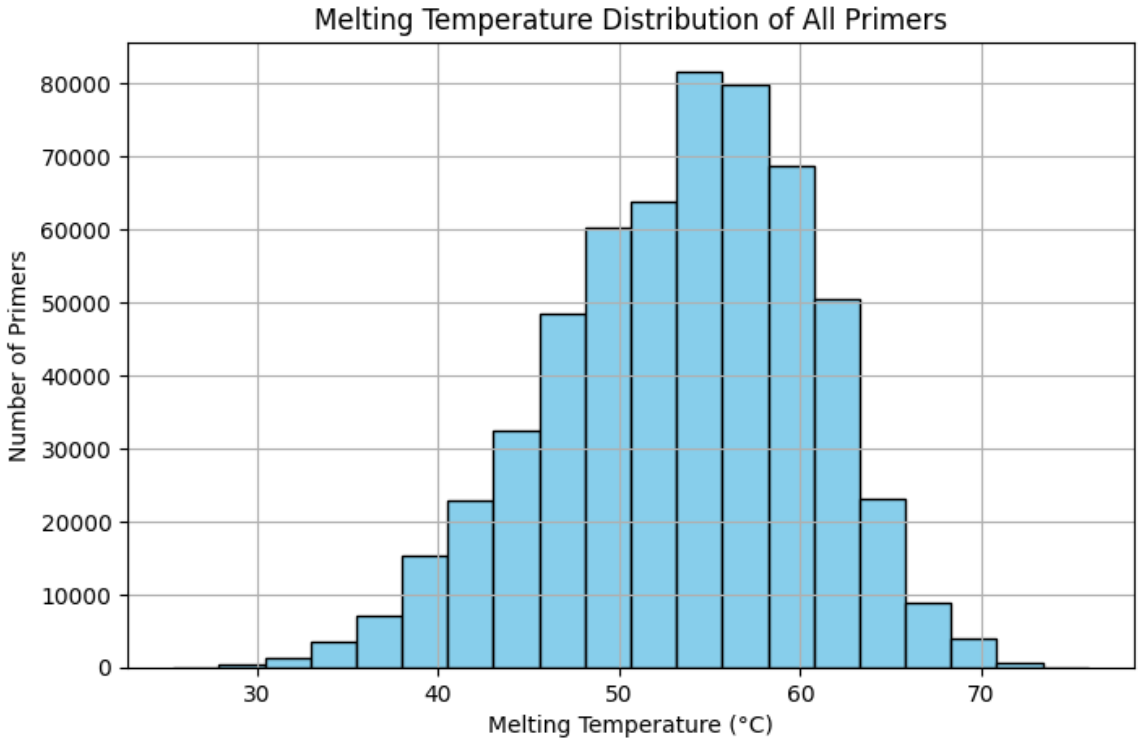


Figure 20. Histogram showing the Melting Temperature (°C) distribution of all iSOP-generated primer sequences.

The majority of iSOP-generated primers fell within the 45-62°C range, with a distinct peak at 53-58°C mark. This coincides with the reaction temperatures used in multiple RT-PCR assays approved by the CDC for emergency use during the 2016 outbreak, which generally ranged from 55-58°C (Altona Diagnostics, 2020; Centers for Disease Control and Prevention, 2016b, 2016c). For the final primer pairs, average temperatures stayed within these ranges as well, varying between 50-55°C and staying within 1°C of each other (**Figure 21**).

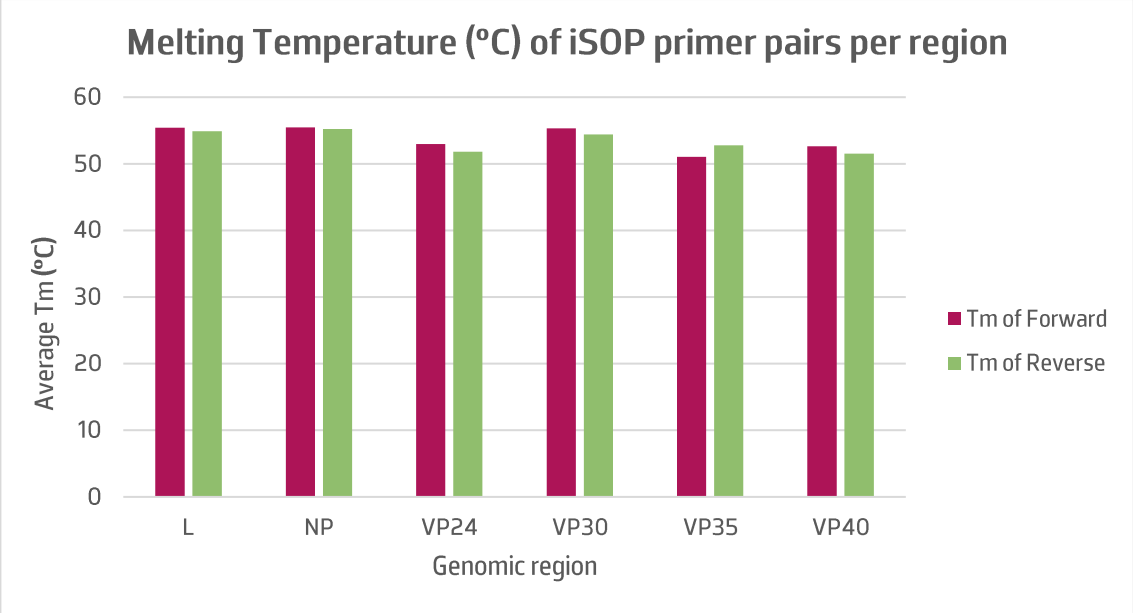


Figure 21. Melting Temperature (°C) average for each primer type (Forward and Reverse) in the iSOP pairs generated by the Combinations script, per genome region.

3.3.4.5 Conservation Scores and Primer combinations

As indicated by the GC Content and Melting Temperature charts for primer pairs, applying a filter to include only primers with Conservation Scores of 99% or higher significantly altered the distribution of targeted genes across the dataset (see **Figure 22**). This stringent filtering was necessary to prevent an unmanageable number of possible primer combinations, but it also resulted in a marked shift in which genes were represented among the selected primers. Consequently, the makeup of the primer set became more focused and selective, emphasizing only those regions with the highest degree of conservation, as reflected in the subsequent figures and analyses.

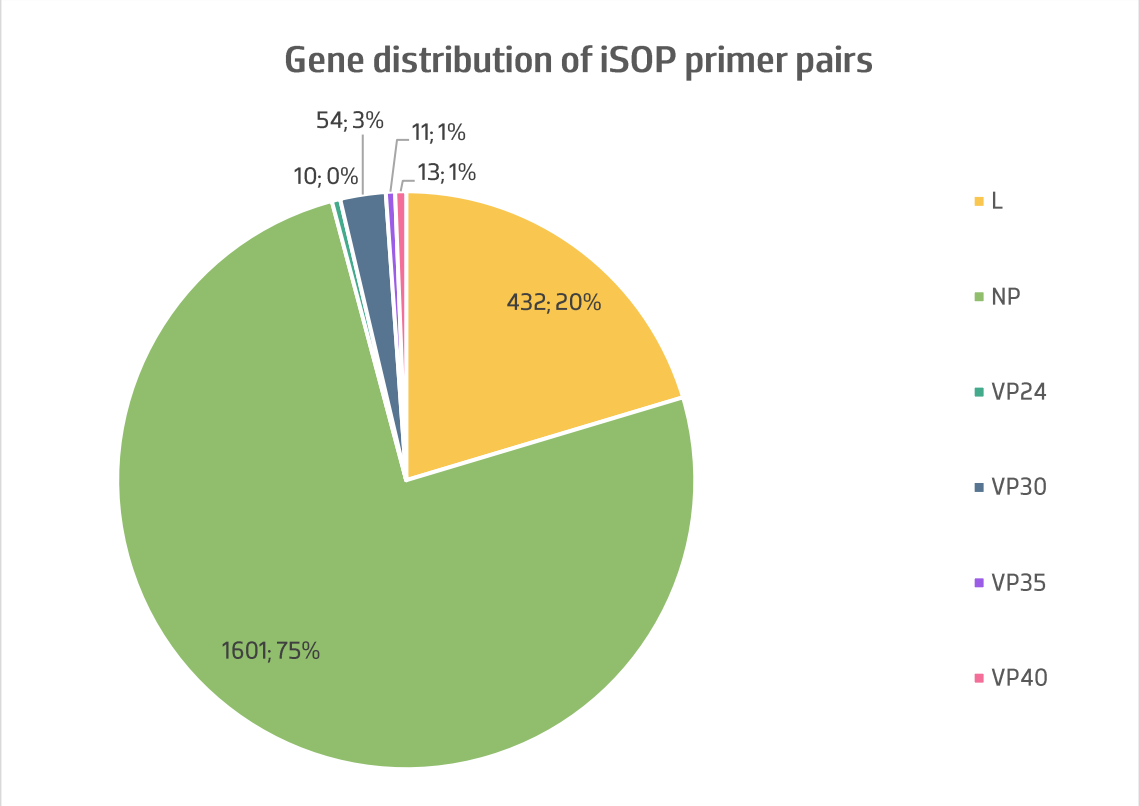


Figure 22. Absolute values and corresponding percentages of iSOP primer pairs, per genome region.

A total of 2121 primer pairs were generated, of which an overwhelming majority (75%) targeted the NP gene. While the L, VP30 and VP35 genes also stayed in the distribution, the complete disappearance of genes like GP, which is widely targeted experimentally, is unexpected. Nevertheless, a similar conclusion can be drawn as in the AROLit analysis: although the GP gene remains relatively well conserved across the dataset, it still exhibits enough sequence variability that primers designed to target this region consistently fail to meet the stringent 99% CS threshold. As a result, these primers are excluded from the final selection, underscoring the challenge of designing highly conserved primers for regions like GP that, while generally stable, harbor enough genetic diversity to fall short of the most rigorous conservation criteria.

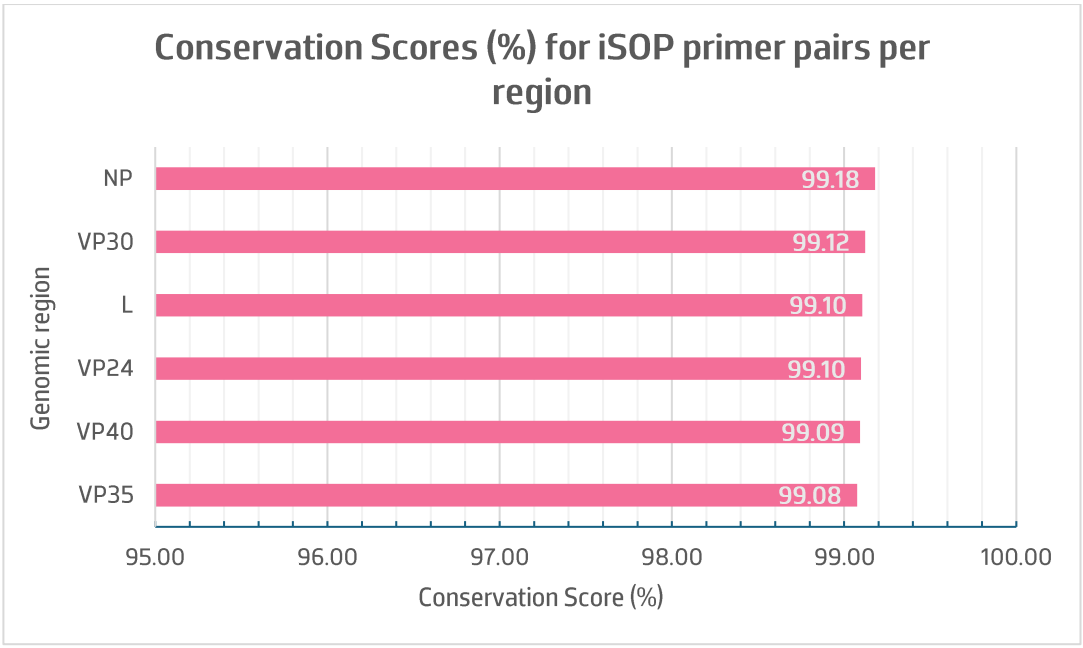


Figure 23. Conservation Score (%) averages for iSOP primer pairs, per genomic region. Due to value similarity, extra labels were added with the exact score average (%).

The filtering step made it so the CS across all regions would inherently be within close range of one another, yet we can still glean some information from **Figure 23**. Primers that fulfill the 99% CS requirement are all similarly conserved, with only a 0.1% difference between the lowest (VP35) and highest (NP) values. Although the differences are small, the fact that NP has a slight advantage over other regions can be due to its lower propensity for accumulating SNPs compared to other genes (Bell et al., 2015). But perhaps a better approach for analyzing scores that are this uniform is by splitting the parameter into its individual elements, PPI and PPI3' % (**Figure 24**).

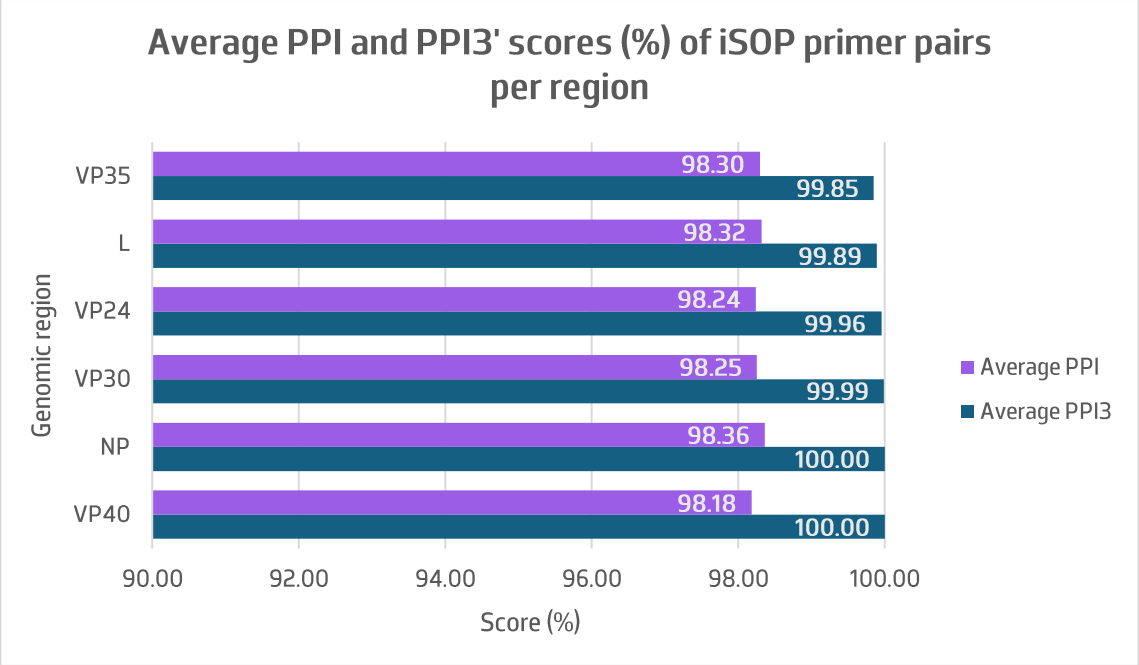


Figure 24. Average PPI and PPI3' scores per genome region. The values correspond to the average scores of both forward and reverse primers for each parameter.

Separating the values that make up CS%, we can now see that what "brings down" the overall score is the PPI, while PPI3' scores approach 100% identity, with NP and VP40 boasting perfect matches across all primers. While both values are extremely important to PCR success, mismatches at the 3' end of a primer are much more critical than those in the 5' end or intermediate regions, due to their potential to disrupt the polymerase's active site (Dieffenbach et al., 1993; Rychlik, 1995; Stadhouders et al., 2010). Perfect PPI3' scores for VP40 and NP primers might shed light on why reputable agencies like the CDC chose these genes as targets for the development of emergency RT-PCR assays, during the most recent Ebola outbreak (Centers for Disease Control and Prevention, 2016b, 2016c).

3.3.4.6 No-Fold%

With PPI and PPI3' scores being established as nearly perfect and uniform across regions, the No-Fold% score differences can consequently be interpreted as more reflective of parameters other than conservation, namely the sequence composition and the thermodynamics behavior of the primers themselves (**Figure 25**). This means that this metric is geared towards intrinsic primer design properties, while CS% is more appropriate for evaluating cross-species coverage.

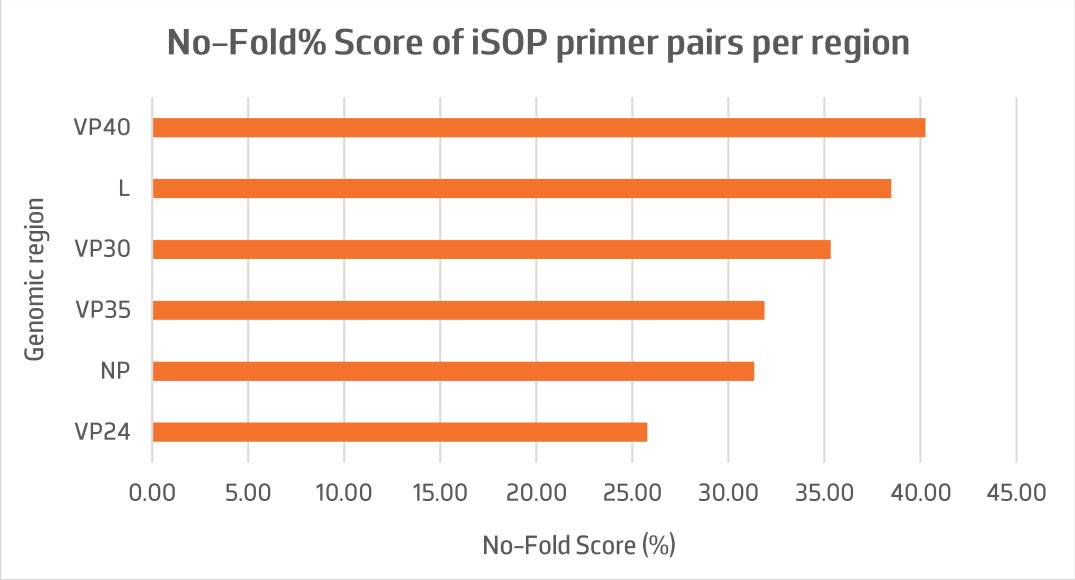


Figure 25. Average of No-Fold% Scores for iSOP primer pairs, per genomic region

Generally speaking, the No-Fold scores for iSOP primers are better than AROLit ones, with VP40 and L pairs yielding the highest average scores of 40 and 39%, respectively. Contrastingly, at only 26%, VP24 primer pairs showed the most risk of hairpin formation and dimerization, despite its excellent conservation score.

If we sort the primers for highest to lowest No-Fold%, however, we get results that make the L gene's earlier dominance reemerge: the first 22 best primer pairs all target this region, with the top 4 pairs all having a No-Fold score of approximately 80.8%. Besides presenting with remarkable stability in comparison to other regions' primers, the best pair fulfills multiple other requirements listed in most Primer Design guidelines (**Table 9**), making it an all-around good candidate for standard PCR experiments (up to 1000 bp amplicon length).

Table 9. Comparison of standard PCR primer design benchmarks with the calculated parameters of the top-performing iSOP primer pair, highlighting alignment with established guidelines and illustrating optimal primer characteristics for reliable PCR amplification.

Generalized benchmarks		Best iSOP primer pair	
		Forward	Reverse
Primer length	18–24 bp	21 bp	21 bp
T_m	50–60°C	51°C	51°C
ΔT_m	≤ 5°C	0°C	
GC Content	40–60%	43%	43%
3'end composition	1–3 G/C in the last 5 nucleotides	2	3
3'end mismatches	None (PPI3' = 100%)	100%	100%
Consecutive repeats of single or dinucleotides	≤ 4 repeats	Yes	Yes
Max Hairpin ΔG	-2 .0 kcal/mol	4.43 kcal/mol	0 kcal/mol
Max Self Dimer ΔG	-5.0 kcal/mol	-1.8 kcal/mol	-2.0 kcal/mol
Max Cross Dimer ΔG	-5.0 kcal/mol	-1.2 kcal/mol	
Amplicon size	200–1000 bp	566 bp	

This comparison highlights the potential of the iSOP algorithm as a quick primer generator that can then be filtered by multiple parameters, depending on the user's needs, for potential primer pairs that, if not already "ready-to-go", can at least serve to cut down on the lengthier steps of primer design.

3.4. ViruScope application and online database

Expanding on the previous section's conclusion, the transformation of AROLit, iSOP and Primer Combinations scripts into ViruScope, a tool readily available for researchers and technicians is not only useful from a practical perspective, but also further boosts the potential and uniqueness of the accompanying database, ViruScopeDB. For example, a researcher can leverage the iSOP and AROLit modules to identify primer candidates well-suited for PCR experimentation. While these initial primers may not be fully optimized for specialized scenarios, such as unique laboratory protocols, they serve as a valuable foundation for further customization. Users have the flexibility to refine sequences, such as by adding a GC clamp at the 3' end, combining different primer options, or extending the 5' end. Importantly,

the ViruScope database is built so that modified primer sequences can be seamlessly fed back into the system for re-scoring and continuous improvement in a transparent, shareable, and reproducible manner.

In practice, ViruScopeDB was successfully designed to be FAIR: easily **findable** through structured indexing and standardized metadata; **accessible** via open, user-friendly interfaces and unrestricted access to the datasets; **interoperable** by adopting common data formats that allow integration with other bioinformatics tools; and **reusable** by providing clear documentation, persistent identifiers, and curated datasets that can be applied in diverse research contexts (Wilkinson et al., 2016).

3.5. Limitations and future perspectives

Perhaps the most glaring limitation faced during the course of the project was the inability to test the data obtained computationally in an experimental setting, more specifically running PCR experiments with the top primer pairs obtained for each virus, similar to what was done in the validation steps of Carneiro et al. (2023) and Pinheiro et al. (2024)'s results. However, unlike the viruses studied in these theses, which target rabbits exclusively, all three of the highlighted viruses exemplified here are highly pathogenic to humans. For instance, Ebolaviruses are universally classified as Risk Group 4 pathogens, which restricts their handling to Biosafety Level 4 (BSL-4) laboratories, none of which are located in Portugal (Lentzos & D. Koblentz, 2021; Shurtleff et al., 2012). In the European Union specifically, SARS-CoV-2 and HIV are considered Risk Group 3, and while BSL-3 facilities exist in the country, specific training in handling pathogenic agents is obligatory for personnel (Directive 2000/54). But while these restrictions meant that direct validation of the tool could not be performed, an argument could be made that, given that they follow the same architecture as the projects done on the Rabbit Hemorrhagic Disease, the results confirming the specificity and sensitivity of AROLit and iSOP primers obtained in both cases point towards the reliability of this new version as well.

Another limitation faced was the suboptimal text extraction from the retrieved PDF files, which hindered the accuracy of the scraping algorithm. In tests performed using the converted files obtained from MinerU 2.5, PDFs with previously null scores had the primers perfectly extracted. Thus, while the current method is still accurate and scalable, we aim to improve on this aspect of the tool by finding or creating better conversion methods.

Regarding future perspectives, the main goal is to fully automate the pipeline. As of the time of writing, the pipeline starts with an alignment file of sequences. The goal is to implement the automatic retrieval of all of the complete genomes for a given organism, and have the tool also perform the MSA, allowing for the workflow to start directly in the app. Moreover, the only step where the user must leave ViruScope and rely on external software is in the retrieval of the full text PDFs, using the metadata file compiled by ViruScope. To eliminate the reliance on any type of extra software, and fully unify the steps, solutions are being investigated, one of which is the direct parsing of the text from the webpage hosting the article. This is theoretically possible due to the already included DOI and PubMed ID fetching tool, which returns a direct URL to the article page, which should host the full text as the retrieval step only considers open access entries. An additional advantage of this approach would be to bypass the conversion step, as the text would be easily found and parsed by accessing the HTML tags in the page's code. The HTML format would also allow for better detection of primers in tables, as cell content is neatly nested and isolated inside tags. On the other hand, drawbacks can potentially include reduced scalability, depending on how this method affects runtime.

Reworking of the current formulas for the Conservation Score is also a strong consideration for the near future, as both literature analysis and experimental data point towards a weighted average of PPI and PPI3', favoring the second, being more informative and correct for evaluating an oligonucleotide's binding success. A potential final overall score, combining the Conservation Scores and the No-Fold%, has also been put forward as a possibility.

Finally, proteomic analyses were initially projected to be part of this thesis, to fully bring together the multi-omics aspect of the original proposal. While abundant literature research was conducted on the topic of these viruses' encoded proteins, and a structure prediction modelling step was initiated, time constraints required for the modelling and molecular dynamics approach to be put on hold. However, future projections include continuing this approach and eventually adding interactive 3D models for the viral proteins in the database.

4. Conclusion

This thesis set out to bridge a critical gap in molecular diagnostics: the lack of a scalable, automated, and cross-viral platform for primer retrieval, generation, validation, and curation. Through the development of ViruScope and its companion resource ViruScopeDB, the project demonstrates that bioinformatics pipelines can go beyond theoretical workflows to provide tangible, practical tools for researchers and public health professionals.

The results obtained with the Ebola case studies validated the pipeline's ability to reproduce experimental best practices, identify highly conserved regions for primer design, and highlight the potential of targeting less common, previously overlooked genes. Importantly, the analyses underscored the complementarity between literature-derived and computationally generated primers, showing how automation can drastically reduce redundancy and accelerate discovery. Additionally, while this work focused on three representative viruses – HIV, Ebola, and SARS-CoV-2 – the framework was designed to be adaptable. This allows expansion to other pathogens, as well as potential integration of new data types and scoring functions. This adaptability positions ViruScope as a sustainable platform for long-term diagnostic preparedness.

All objectives outlined for this project were successfully achieved. The developed tool enabled automated mining and analysis of viral sequences, literature-based extraction of oligonucleotides, *in silico* primer generation, and validation against genomic alignments. Key features, including automated primer pairing, benchmark scoring of primer parameters, and structured export of results, were fully implemented, ensuring a reliable and user-friendly pipeline. Furthermore, the creation of ViruScopeDB merged and expanded upon previous databases in line with FAIR principles, delivering an open-access, cross-viral repository. Finally, in testing the central hypothesis, we demonstrated that automation and integration of viral genomic and literature-derived data do indeed enhance the efficiency, scalability, and reliability of diagnostic primer design. Together, these outcomes demonstrate that both the tool and database fulfill the overarching aim of providing scalable, automated, and broadly applicable resources for viral diagnostics and preparedness.

In conclusion, this thesis contributes both a methodological advance and a practical resource. By automating primer mining, validation, and curation, ViruScope and ViruScopeDB strengthen the global

capacity to respond to viral outbreaks more rapidly and reliably. As viral evolution continues to challenge diagnostic tools, initiatives such as this highlight the essential role of bioinformatics in building resilient, scalable, and open-access infrastructures for public health.

5. Contributions

Over the course of the project, presentations of the findings were made in the form of three posters and one oral communication, at different scientific conferences focused on bioinformatics and evolutionary biology:

- **XX ENBE Annual Meeting of the Portuguese Association for Evolutionary Biology (2024):** *"A Multi-Omics and Primer Database for Virus Identification: Focus on HIV, Ebola, and SARS-CoV-2"* – poster (DOI: 10.13140/RG.2.2.31826.47041);
- **Bioinformatics Open Days XIV (2025):** *"Comprehensive multi-omics database for highly infectious viruses: a focus on HIV, Ebola and SARS-CoV-2"* – poster (DOI: 10.13140/RG.2.2.21760.14087);
- **Young Researchers Meeting of the U.Porto – IJUP (2025):** *"ViruScopeDB: a comprehensive multi-omics database for highly infectious viruses"* – poster (DOI: 10.13140/RG.2.2.11748.49282);
- **Second Symposium on Biostatistics and Bioinformatics Applied to Health – SBBAH (2025):** *"ViruScopeDB: a comprehensive multi-omics database for highly infectious viruses"* – oral communication (DOI: 10.13140/RG.2.2.19746.34246).

Additionally, as a result of the scope of my project, I was invited to teach a class on Data mining applied to health sciences – where I shared my work as an example of the potential of bioinformatics pipelines to process healthcare data – at the School of Health Sciences – Polytechnic of Porto, for first year Master's students of Biostatistics and Bioinformatics Applied to Health.

These presentations served to both disseminate my findings and promote the results of my work, as well as engage with and learn from other members of the scientific community, establish potential future collaborations and bring to the forefront the importance of proactivity in the face of a world consistently under the threat of emerging pathogens.

The ViruScope project's results, including the scripts, analytics and other publicly available resources can be found at:

- ViruScopeDB: https://viruscope.jc-biotechaiteam.com/wp
- ViruScope tool repositories:
    - https://github.com/anasfplima/ViruScope
    - https://github.com/anasfplima/ViruScope-GUI
- AROLit and iSOP datasets for Ebola: https://doi.org/10.5281/zenodo.17227856

Bibliographic references

Akoi Boré, J., Timothy, J. W. S., Tipton, T., Kekoura, I., Hall, Y., Hood, G., Longet, S., Fornace, K., Lucien, M. S., Fehling, S. K., Koivogui, B. K., Coggins, S. A., Laing, E. D., Broder, C. C., Magassouba, N. F., Strecker, T., Rossman, J., Konde, K., & Carroll, M. W. (2024). Serological evidence of zoonotic filovirus exposure among bushmeat hunters in Guinea. *Nature Communications*, *15*(1), 4171. https://doi.org/10.1038/s41467-024-48587-5

Alexander, K. A., Sanderson, C. E., Marathe, M., Lewis, B. L., Rivers, C. M., Shaman, J., Drake, J. M., Lofgren, E., Dato, V. M., Eisenberg, M. C., & Eubank, S. (2015). What Factors Might Have Led to the Emergence of Ebola in West Africa? *PLoS Neglected Tropical Diseases*, *9*(6), e0003652. https://doi.org/10.1371/journal.pntd.0003652

Altona Diagnostics. (2020, May). *RealStar® Ebolavirus RT-PCR Kit 1.0*. Altona Diagnostics.

Applied Biosystems. (2005). *Real-Time PCR systems. Applied Biosystems 7900HT Fast Real-Time PCR System and 7300/7500 Real-Time PCR Systems. Chemistry Guide*. Applied Biosystems.

Arrildt, K. T., Joseph, S. B., & Swanstrom, R. (2012). The HIV-1 Env Protein: A Coat of Many Colors. *Current HIV/AIDS Reports*, *9*(1), 52–63. https://doi.org/10.1007/s11904-011-0107-3

Behind The Bench Staff. (2019, September 25). PCR Primer Design Tips. *Behind the Bench*. https://www.thermofisher.com/blog/behindthebench/pcr-primer-design-tips/

Bell, A., Lewandowski, K., Myers, R., Wooldridge, D., Aarons, E., Simpson, A., Vipond, R., Jacobs, M., Gharbia, S., & Zambon, M. (2015). Genome sequence analysis of Ebola virus in clinical samples from three British healthcare workers, August 2014 to March 2015. *Eurosurveillance*, *20*(20). https://doi.org/10.2807/1560-7917.ES2015.20.20.21131

Benchling. (n.d.). *Guide to Primer Design for PCR*. Retrieved September 20, 2025, from https://www.benchling.com/primer-design-for-pcr

Bettini, A., Lapa, D., & Garbuglia, A. R. (2023). Diagnostics of Ebola virus. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.1123024

Bio-Rad Laboratories. (2006). *Real-Time PCR Applications Guide*. Bio-Rad Laboratories.

Bustin, S., & Huggett, J. (2017). qPCR primer design revisited. *Biomolecular Detection and Quantification*, *14*, 19–28. https://doi.org/10.1016/j.bdq.2017.11.001

Carneiro, F., Carneiro, J., Abrantes, J., & Lopes, A. M. (2023). *A computational method to build an RHDV primer database: Literature and in silico approaches*. Faculty of Sciences – University of Porto.

Carneiro, J., Gomes, C., Couto, C., & Pereira, F. (2020). *CoV2ID: Detection and Therapeutics Oligo Database for SARS-CoV-2* (p. 2020.04.19.048991). bioRxiv. https://doi.org/10.1101/2020.04.19.048991

Carneiro, J., & Pereira, F. (2016a). EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. *PLOS Neglected Tropical Diseases*, *10*(7), e0004757. https://doi.org/10.1371/journal.pntd.0004757

Carneiro, J., & Pereira, F. (2016b). EbolaID: An Online Database of Informative Genomic Regions for Ebola Identification and Treatment. *PLOS Neglected Tropical Diseases*, *10*(7), e0004757. https://doi.org/10.1371/journal.pntd.0004757

Carneiro, J., Resende, A., & Pereira, F. (2017). The HIV oligonucleotide database (HIVoligoDB). *Database*, *2017*, bax005. https://doi.org/10.1093/database/bax005

Carroll, M. W., Matthews, D. A., Hiscox, J. A., Elmore, M. J., Pollakis, G., Rambaut, A., Hewson, R., García-Dorival, I., Bore, J. A., Koundouno, R., Abdellati, S., Afrough, B., Aiyepada, J., Akhilomen, P., Asogun, D., Atkinson, B., Badusche, M., Bah, A., Bate, S., … Günther, S. (2015). Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature*, *524*(7563), 97–101. https://doi.org/10.1038/nature14594

Carter, M., & Shieh, J. (2015). Chapter 11—Gene Delivery Strategies. In M. Carter & J. Shieh (Eds.), *Guide to Research Techniques in Neuroscience (Second Edition)* (pp. 239–252). Academic Press. https://doi.org/10.1016/B978-0-12-800511-8.00011-3

Centers for Disease Control and Prevention. (2016a). *Cost of the Ebola Epidemic*. Centers for Disease Control and Prevention. https://stacks.cdc.gov/view/cdc/40043/cdc_40043_DS1.pdf

Centers for Disease Control and Prevention. (2016b, January). *Ebola Virus NP Real-Time RT-PCR Assay*. Centers for Disease Control and Prevention. https://www.fda.gov/media/91097/download

Centers for Disease Control and Prevention. (2016c, January). *Ebola Virus VP40 Real-Time RT-PCR Assay*. Centers for Disease Control and Prevention. https://www.fda.gov/media/91142/download

Chen, J., Zhou, T., Zhang, Y., Luo, S., Chen, H., Chen, D., Li, C., & Li, W. (2022). The reservoir of latent HIV. *Frontiers in Cellular and Infection Microbiology*, *12*, 945956. https://doi.org/10.3389/fcimb.2022.945956

Cohen, J. P., Anupindi, V. R., Doshi, R., Yeaw, J., Zhou, X., Christoph, M. J., Chen, M., Chaudhari, P., Trom, C., & Zachry, W. (2025). Estimation of Lifetime Costs Among Insured Persons with HIV in the United States. *PharmacoEconomics - Open*. https://doi.org/10.1007/s41669-025-00584-0

Craigie, R., & Bushman, F. D. (2012). HIV DNA Integration. *Cold Spring Harbor Perspectives in Medicine*, *2*(7), a006890. https://doi.org/10.1101/cshperspect.a006890

Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J., & Sanjuán, R. (2015). Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biology*, *13*(9), e1002251. https://doi.org/10.1371/journal.pbio.1002251

Cutler, D. M., & Summers, L. H. (2020). The COVID-19 Pandemic and the $16 Trillion Virus. *JAMA*, *324*(15), 1495–1496. https://doi.org/10.1001/jama.2020.19759

Debode, F., Marien, A., Jansen, É., Bragard, C., & Berben, G. (2017). The influence of amplicon length on real-time PCR results. *Biotechnology, Agronomy, Society and Environment. BASE*, *21*(1), 3-11. https://doi.org/10.25518/1780-4507.13461

Denison, M. R., Graham, R. L., Donaldson, E. F., Eckerle, L. D., & Baric, R. S. (2011). Coronaviruses. *RNA Biology*, *8*(2), 270–279. https://doi.org/10.4161/rna.8.2.15013

Dieffenbach, C. W., Lowe, T. M., & Dveksler, G. S. (1993). General concepts for PCR primer design. *Genome Research*, *3*(3), S30–S37. https://doi.org/10.1101/gr.3.3.S30

Diehl, W. E., Lin, A. E., Grubaugh, N. D., Carvalho, L. M., Kim, K., Kyawe, P. P., McCauley, S. M., Donnard, E., Kucukural, A., McDonel, P., Schaffner, S. F., Garber, M., Rambaut, A., Andersen, K. G., Sabeti, P. C., & Luban, J. (2016). Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic. *Cell*, *167*(4), 1088-1098.e6. https://doi.org/10.1016/j.cell.2016.10.014

Directive 2000/54/EC of the European Parliament and of the Council of 18 September 2000 on the Protection of Workers from Risks Related to Exposure to Biological Agents at Work (Seventh Individual Directive within the Meaning of Article 16(1) of Directive 89/391/EEC) (2020). http://data.europa.eu/eli/dir/2000/54/2020-06-24/eng

Dowdle, W. R. (1998). The principles of disease elimination and eradication. *Bulletin of the World Health Organization*, *76*(Suppl 2), 22–25. https://pubmed.ncbi.nlm.nih.gov/10063669/

Duchon, A., & Hu, W.-S. (2024). HIV-1 RNA genome packaging: It's G-rated. *mBio*, *15*(4), e00861-23. https://doi.org/10.1128/mbio.00861-23

Ebertz, A. (2022, September 5). Primer Design Guide. *The DNA Universe BLOG*. https://the-dna-universe.com/2022/09/05/primer-design-guide-the-top-5-factors-to-consider-for-optimum-performance/

Epidemic, Endemic, Pandemic: What are the Differences? (2021, February 19). Columbia University Mailman School of Public Health. https://www.publichealth.columbia.edu/news/epidemic-endemic-pandemic-what-are-differences

Garibyan, L., & Avashia, N. (2013). Research Techniques Made Simple: Polymerase Chain Reaction (PCR). *The Journal of Investigative Dermatology*, *133*(3), e6. https://doi.org/10.1038/jid.2013.1

Geneious Prime. (n.d.-a). *Practice Primer Design*. Retrieved September 16, 2025, from https://www.geneious.com/tutorials/primer-design

Geneious Prime (Version 2025.1). (n.d.-b). [Computer software]. Biomatters. https://www.geneious.com

Gribble, J., Stevens, L. J., Agostini, M. L., Anderson-Daniels, J., Chappell, J. D., Lu, X., Pruijssers, A. J., Routh, A. L., & Denison, M. R. (2021). The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathogens*, *17*(1), e1009226. https://doi.org/10.1371/journal.ppat.1009226

Haileamlak, A. (2021). The impact of COVID-19 on health and health systems. *Ethiopian Journal of Health Sciences*, *31*(6), 1073–1074. https://doi.org/10.4314/ejhs.v31i6.1

Hall-Wheeler, C. (2015, April 1). *Primer Design Considerations*. University of Nevada. https://www.unlv.edu/genomics/equipment-services/primer-design-considerations

Hao, Y., Wang, Y., Wang, M., Zhou, L., Shi, J., Cao, J., & Wang, D. (2022). The origins of COVID-19 pandemic: A brief overview. *Transboundary and Emerging Diseases*, 10.1111/tbed.14732. https://doi.org/10.1111/tbed.14732

Hendy, M., Kaufman, S., & Ponga, M. (2021). Molecular strategies for antibody binding and escape of SARS-CoV-2 and its mutations. *Scientific Reports*, *11*, 21735. https://doi.org/10.1038/s41598-021-01081-0

Hoenen, T., Groseth, A., & Feldmann, H. (2012). Current Ebola vaccines. *Expert Opinion on Biological Therapy*, *12*(7), 859–872. https://doi.org/10.1517/14712598.2012.685152

Humans, I. W. G. on the E. of C. R. to. (2012). HUMAN IMMUNODEFICIENCY VIRUS-1. In *Biological Agents*. International Agency for Research on Cancer. https://www.ncbi.nlm.nih.gov/books/NBK304351/

Jääskeläinen, A. J., Sironen, T., Diagne, C. T., Diagne, M. M., Faye, M., Faye, O., Faye, O., Hewson, R., Mölsä, M., Weidmann, M. W., Watson, R., Sall, A. A., & Vapalahti, O. (2019). Development, validation and clinical evaluation of a broad-range pan-filovirus RT-qPCR. *Journal of Clinical Virology*, *114*, 26–31. https://doi.org/10.1016/j.jcv.2019.03.010

Jääskeläinen, A. J., Sironen, T., Kaloinen, M., Kakkola, L., Julkunen, I., Hewson, R., Weidmann, M. W., Mirazimi, A., Watson, R., & Vapalahti, O. (2020). Comparison of Zaire ebolavirus realtime RT-PCRs targeting the nucleoprotein gene. *Journal of Virological Methods*, *284*, 113941. https://doi.org/10.1016/j.jviromet.2020.113941

Jackson, C. B., Farzan, M., Chen, B., & Choe, H. (2022). Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology*, *23*(1), 3–20. https://doi.org/10.1038/s41580-021-00418-x

Jun, S.-R., Leuze, M. R., Nookaew, I., Uberbacher, E. C., Land, M., Zhang, Q., Wanchai, V., Chai, J., Nielsen, M., Trolle, T., Lund, O., Buzard, G. S., Pedersen, T. D., Wassenaar, T. M., & Ussery, D. W. (2015). Ebolavirus comparative genomics. *FEMS Microbiology Reviews*, *39*(5), 764–778. https://doi.org/10.1093/femsre/fuv031

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kielian, M., & Rey, F. A. (2006). Virus membrane-fusion proteins: More than one way to make a hairpin. *Nature Reviews Microbiology*, *4*(1), 67–76. https://doi.org/10.1038/nrmicro1326

Koehler, J. W., Stefan, C. P., Hall, A. T., Delp, K. L., O'Hearn, A. E., Taylor-Howell, C. L., Wauquier, N., Schoepp, R. J., & Minogue, T. D. (2023). Sequence optimized diagnostic assay for Ebola virus detection. *Scientific Reports*, *13*(1), 18840. https://doi.org/10.1038/s41598-023-29390-6

Lentivirus Fact Sheet. (n.d.). Retrieved September 23, 2025, from https://ehs.stanford.edu/reference/lentivirus-fact-sheet

Lentzos, F., & D. Koblentz, G. (2021, May). *Mapping Maximum Biological Containment Labs Globally*. King's College London. Retrieved September 16, 2025, from https://www.globalbiolabs.org/

Madadelahi, M., Agarwal, R., Martinez-Chapa, S. O., & Madou, M. J. (2024). A roadmap to high-speed polymerase chain reaction (PCR): COVID-19 as a technology accelerator. *Biosensors and Bioelectronics*, *246*, 115830. https://doi.org/10.1016/j.bios.2023.115830

Mahale, K. N., & Patole, M. S. (2015). The crux and crust of ebolavirus: Analysis of genome sequences and glycoprotein gene. *Biochemical and Biophysical Research Communications*, *463*(4), 756–761. https://doi.org/10.1016/j.bbrc.2015.06.008

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Gavrilov, D., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D., Ortiz-Ospina, E., & Roser, M. (2020). Coronavirus (COVID-19) Deaths. *Our World in Data*. Retrieved September 16, 2025, from https://ourworldindata.org/covid-deaths

Moore, M. D., & Hu, W.-S. (2009). HIV-1 RNA Dimerization: It Takes Two to Tango. *AIDS Reviews*, *11*(2), 91–102. https://pmc.ncbi.nlm.nih.gov/articles/PMC3056336/

Naqvi, A. A. T., Fatima, K., Mohammad, T., Fatima, U., Singh, I. K., Singh, A., Atif, S. M., Hariprasad, G., Hasan, G. M., & Hassan, Md. I. (2020). Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta. Molecular Basis of Disease*, *1866*(10), 165878. https://doi.org/10.1016/j.bbadis.2020.165878

Onodera, K., & Melcher, U. (2002). VirOligo: A database of virus-specific oligonucleotides. *Nucleic Acids Research*, *30*(1), 203–204. https://doi.org/10.1093/nar/30.1.203

Ossola, A. (2020, March 25). *Here are the coronavirus testing materials that are in short supply in the US*. Quartz. https://qz.com/1822596/all-the-coronavirus-test-materials-in-short-supply-in-the-us

Pinheiro, M. C., Carneiro, J., Lopes, A. M., & Pratas, D. (2024). *Genomic Diversity and Zoonotic Potential of Hepatitis E Virus in European Rabbits: Implications for Diagnostic and Therapeutic Approaches*. Faculty of Sciences – University of Porto.

Premier Biosoft. (2025). *Primer Design Guide for PCR: Learn Designing Primers for PCR*. https://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html

Public Health England. (2020). *PHE novel coronavirus diagnostic test rolled out across UK*. GOV.UK. https://www.gov.uk/government/news/phe-novel-coronavirus-diagnostic-test-rolled-out-across-uk

Rhodes, T. D., Nikolaitchik, O., Chen, J., Powell, D., & Hu, W.-S. (2005). Genetic Recombination of Human Immunodeficiency Virus Type 1 in One Round of Viral Replication: Effects of Genetic Distance, Target Cells, Accessory Genes, and Lack of High Negative Interference in Crossover Events. *Journal of Virology*, *79*(3), 1666–1677. https://doi.org/10.1128/JVI.79.3.1666-1677.2005

Roberts, J. D., Bebenek, K., & Kunkel, T. A. (1988). The Accuracy of Reverse Transcriptase from HIV-1. *Science*, *242*(4882), 1171–1173. https://doi.org/10.1126/science.2460925

Rychlik, W. (1995). Selection of primers for polymerase chain reaction. *Molecular Biotechnology*, *3*(2), 129–134. https://doi.org/10.1007/BF02789108

Sanchez, A., Kiley, M. P., Holloway, B. P., & Auperin, D. D. (1993). Sequence analysis of the Ebola virus genome: Organization, genetic elements, and comparison with the genome of Marburg virus. *Virus Research*, *29*(3), 215–240. https://doi.org/10.1016/0168-1702(93)90063-S

Sanchez, A., & Rollin, P. E. (2005). Complete genome sequence of an Ebola virus (Sudan species) responsible for a 2000 outbreak of human disease in Uganda. *Virus Research*, *113*(1), 16–25. https://doi.org/10.1016/j.virusres.2005.03.028

Shahhosseini, N., Babuadze, G. (Giorgi), Wong, G., & Kobinger, G. P. (2021). Mutation Signatures and In Silico Docking of Novel SARS-CoV-2 Variants of Concern. *Microorganisms*, *9*(5), 926. https://doi.org/10.3390/microorganisms9050926

Sharp, P. M., & Hahn, B. H. (2011). Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine*, *1*(1), a006841. https://doi.org/10.1101/cshperspect.a006841

Shurtleff, A. C., Garza, N., Lackemeyer, M., Carrion, R. J., Griffiths, A., Patterson, J., Edwin, S. S., & Bavari, S. (2012). The Impact of Regulations, Safety Considerations and Physical Limitations on Research Progress at Maximum Biocontainment. *Viruses*, *4*(12), 3932–3951. https://doi.org/10.3390/v4123932

Sigma Aldrich. (n.d.). *OligoArchitect Online—Glossary of Parameters*. Retrieved September 20, 2025, from https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/marketing/global/documents/200/845/oligo-architect-glossary-br3011en-mk.pdf

Silva, J. M., Pinho, A. J., & Pratas, D. (2024). AltaiR: A C toolkit for alignment-free and temporal analysis of multi-FASTA data. *GigaScience*, *13*, giae086. https://doi.org/10.1093/gigascience/giae086

Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H. M., & Schutten, M. (2010). The Effect of Primer-Template Mismatches on the Detection and Quantification of Nucleic Acids Using the 5′ Nuclease Assay. *The Journal of Molecular Diagnostics : JMD*, *12*(1), 109–117. https://doi.org/10.2353/jmoldx.2010.090035

Swetha, R. G., Ramaiah, S., Anbarasu, A., & Sekar, K. (2016). Ebolavirus Database: Gene and Protein Information Resource for Ebolaviruses. *Advances in Bioinformatics*, *2016*(1), 1673284. https://doi.org/10.1155/2016/1673284

Temple-Raston, D. (2020, November 6). CDC Report: Officials Knew Coronavirus Test Was Flawed But Released It Anyway. *NPR*. https://www.npr.org/2020/11/06/929078678/cdc-report-officials-knew-coronavirus-test-was-flawed-but-released-it-anyway

The COVID Decade: Understanding the long-term societal impacts of COVID-19. (2021). The British Academy. https://www.thebritishacademy.ac.uk/publications/covid-decade-understanding-the-long-term-societal-impacts-of-covid-19/

Trombley, A. R., Wachter, L., Garrison, J., Buckley-Beason, V. A., Jahrling, J., Hensley, L. E., Schoepp, R. J., Norwood, D. A., Goba, A., Fair, J. N., & Kulesh, D. A. (2010). Comprehensive Panel of Real-Time TaqMan™ Polymerase Chain Reaction Assays for Detection and Absolute Quantification of Filoviruses, Arenaviruses, and New World Hantaviruses. *The American Society of Tropical Medicine and Hygiene*, *82*(5), 954–960. https://doi.org/10.4269/ajtmh.2010.09-0636

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wong, G., He, S., Leung, A., Cao, W., Bi, Y., Zhang, Z., Zhu, W., Wang, L., Zhao, Y., Cheng, K., Liu, D., Liu, W., Kobasa, D., Gao, G. F., & Qiu, X. (2018). Naturally Occurring Single Mutations in Ebola Virus Observably Impact Infectivity. *Journal of Virology*, *93*(1), e01098-18. https://doi.org/10.1128/JVI.01098-18

Woo, P. C. Y., Huang, Y., Lau, S. K. P., & Yuen, K.-Y. (2010). Coronavirus Genomics and Bioinformatics Analysis. *Viruses*, *2*(8), 1804–1820. https://doi.org/10.3390/v2081803

World Health Organization. (n.d.). *Ebola outbreak 2014-2016—West Africa*. Retrieved September 22, 2025, from https://www.who.int/emergencies/situations/ebola-outbreak-2014-2016-West-Africa

World Health Organization. (2024, August 7). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

World Health Organization. (2025a). *HIV data and statistics*. https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics

World Health Organization. (2025b, April 24). *Ebola disease*. https://www.who.int/news-room/fact-sheets/detail/ebola-disease

World Health Organization, International Labour Organization, Food and Agriculture Organization, & International Fund for Agricultural Development. (2020, October 13). *Impact of COVID-19 on people's livelihoods, their health and our food systems* [Statement]. https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems

Yang, M., Ke, Y., Zhang, W., Liu, C., Yang, R., & Chen, Z. (2017). RT-PCR using glycoprotein target is more sensitive for the detection of Ebola virus in clinical samples. *Diagnostic Microbiology and Infectious Disease*, *87*(3), 235–237. https://doi.org/10.1016/j.diagmicrobio.2016.11.001

Yeo, J. Y., Goh, G.-R., Su, C. T.-T., & Gan, S. K.-E. (2020a). The Determination of HIV-1 RT Mutation Rate, Its Possible Allosteric Effects, and Its Implications on Drug Resistance. *Viruses*, *12*(3), 297. https://doi.org/10.3390/v12030297

Yeo, J. Y., Goh, G.-R., Su, C. T.-T., & Gan, S. K.-E. (2020b). The Determination of HIV-1 RT Mutation Rate, Its Possible Allosteric Effects, and Its Implications on Drug Resistance. *Viruses*, *12*(3), 297. https://doi.org/10.3390/v12030297

Yu, F., Wen, Y., Wang, J., Gong, Y., Feng, K., Ye, R., Jiang, Y., Zhao, Q., Pan, P., Wu, H., Duan, S., Su, B., & Qiu, M. (2018). The Transmission and Evolution of HIV-1 Quasispecies within One Couple: A Follow-up Study based on Next-Generation Sequencing. *Scientific Reports*, *8*(1), 1404. https://doi.org/10.1038/s41598-018-19783-3

Zaire ebolavirus isolate Ebola virus/H.sapiens-tc/COD/1976/Yambuku-Mayinga, complete genome (10313991). (2018). [Dataset]. NCBI Nucleotide Database. http://www.ncbi.nlm.nih.gov/nuccore/NC_002549.1

Zhao, J., Guo, S., Yi, D., Li, Q., Ma, L., Zhang, Y., Wang, J., Li, X., Guo, F., Lin, R., Liang, C., Liu, Z., & Cen, S. (2021). A cell-based assay to discover inhibitors of SARS-CoV-2 RNA dependent RNA polymerase. *Antiviral Research*, *190*, 105078. https://doi.org/10.1016/j.antiviral.2021.105078

P.PORTO

ESCOLA SUPERIOR DE SAÚDE

M

MESTRADO

BIOESTATÍSTICA E BIOINFORMÁTICA APLICADAS À SAÚDE