ACS Infectious Diseases

pubs.acs.org/journal/aidcbc

Article

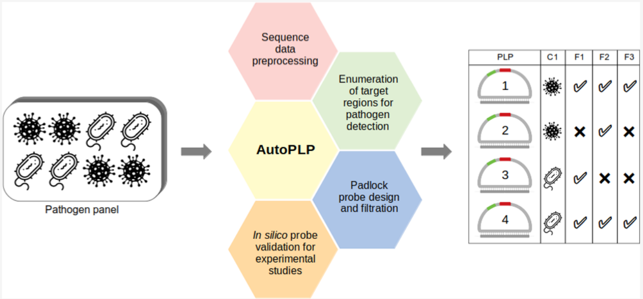
AutoPLP: A Padlock Probe Design Pipeline for Zoonotic Pathogens

Sowmya Ramaswamy Krishnan, Ruben R. G. Soares, Narayanan Madaboosi,* and M. Michael Gromiha*

Cite This: *ACS Infect. Dis.* 2023, 9, 459−469

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information


Pathogen panel
Sequence data preprocessing
Enumeration of target regions for pathogen detection
AutoPLP
Padlock probe design and filtration
In silico probe validation for experimental studies
PLP
C1
F1
F2
F3
1
2
3
4

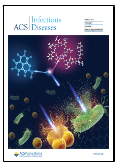
ABSTRACT: Emergence of novel zoonotic infections among the human population has increased the burden on global healthcare systems to curb their spread. To meet the evolutionary agility of pathogens, it is essential to revamp the existing diagnostic methods for early detection and characterization of the pathogens at the molecular level. Padlock probes (PLPs), which can leverage the power of isothermal nucleic acid amplification techniques (NAAT) such as rolling circle amplification (RCA), are known for their high sensitivity and specificity in detecting a diverse pathogen panel of interest. However, due to the complexity involved in deciding the target regions for PLP design and the need for optimization of multiple experimental parameters, the applicability of RCA has been limited in point-of-care testing for pathogen detection. To address this gap, we have developed a novel and integrated PLP design pipeline named AutoPLP, which can automate the probe design process for a diverse pathogen panel of interest. The pipeline is composed of three modules which can perform sequence data curation, multiple sequence alignment, conservation analysis, filtration based on experimental parameters (T_m, GC content, and secondary structure formation), and *in silico* probe validation via potential cross-hybridization check with host genome. The modules can also take into account the backbone and restriction site information, appropriate combinations of which are incorporated along with the probe arms to design a complete probe sequence. The potential applications of AutoPLP are showcased through the design of PLPs for the detection of rabies virus and drug-resistant strains of *Mycobacterium tuberculosis.*

KEYWORDS: *autoPLP, padlock probe, RCA, zoonotic pathogens, rabies, Mycobacterium tuberculosis*

The spread of infectious diseases has recently increased at an alarming pace, with millions of lives being afflicted everyday.[1] Infections such as the recently alarming COVID-19 have led to heavy socioeconomic burdens, thus directly affecting the healthcare systems.[2] Furthermore, the complex and multifaceted nature of infectious diseases challenges our ability to curb their spread, thereby providing the opportunity for pathogens to further enhance their evolutionary fitness in the environment.[3] The resultant noticeable increase in pathogen load has alongside supported the evolution of several antimicrobial resistance (AMR) strains.[4] With the looming threat of AMR, dysfunctional treatment regimens involving first- and second-line antibiotics demand attention with adapting bacterial strains; likewise, the increase in the number of emerging and reemerging strains of hypervariable viruses at a global scale equally burdens the healthcare systems.[5,6] Hence, it is highly necessary to find alternate solutions to diagnose and tackle infectious diseases in an integrated and rapid manner. Especially, natural enzymatic adaptations observed in RNA viruses such as the error-prone polymerase functionality confer high sequence diversity and capability to bypass the species barrier, thereby disrupting not only the existing diagnostic strategies to detect them but also the available therapeutic options to treat such viral infections.[7] Further, the increase in interactions between humans, animals, and pathogens, resulting from increased migration and travel trends, has inevitably led to an increased risk for novel zoonotic infections, requiring cross-species surveillance and rapid control of spread among both human and animal populations.[8] Hence, rapid diagnostic techniques with high sensitivity and specificity are required for early detection of zoonoses caused by AMR strains of pathogenic bacteria and hypervariable strains of viruses, which can aid better surveillance, prognosis, and response.

The era of biotechnology revolution has gifted humanity with an array of bioassay methodologies and advanced readout platforms to precisely enhance the molecular identification of pathogens in both temporal and spatial scales.[9] Popular diagnostic methods for pathogen detection can be divided into several subcategories including microbiological assays targeting the phenotype, serological antigen/antibody detection assays, and nucleic acid amplification tests (NAATs) directly targeting the genotype. Preliminary characterization of the micro-organism of interest is often accomplished through

Received: August 23, 2022
Published: February 15, 2023



https://doi.org/10.1021/acsinfecdis.2c00436

microbiological techniques involving culturing of the pathogen from the infected specimen through selective and nonselective enrichment methods, followed by initial identification of the pathogen from the culture based on selective staining procedures, and subsequent biochemical confirmation of the detected pathogen using spectroscopic or other relevant methods.[10] The immune response profile of the infected individual is further studied using serological assays involving detection of highly specific antigen-antibody interactions. These methods include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), immunofluorescence assay (IFA), latex agglutination assay, and hemagglutination assay. Serological assays designed based on highly pathogen-specific surface antigens can help pinpoint the pathogens of interest responsible for the zoonotic infections, based on the corresponding antibody generated by the human immune response pathways.[11]

NAATs are potential alternatives to microbiological and serological assays due to their relatively faster mode of detection and high sensitivity, which are desirable for constantly monitoring the continually evolving pathogens.[9,12,13] The most popular NAATs include polymerase chain reaction (PCR), rolling circle amplification (RCA), loop-mediated isothermal amplification (LAMP), recombinase polymerase amplification (RPA), strand displacement amplification (SDA), and nucleic acid sequence-based amplification (NASBA). Among these methods, RCA, LAMP, RPA, SDA, and NASBA can be classified as isothermal NAATs, as they can work under a constant temperature, thereby becoming relevant for applications at the point-of-care (PoC) settings. However, PCR depends on a carefully designed temperature ramping protocol to enable the nucleic acid of interest to go through the three phases of transformation: denaturation, annealing, and extension.[9] Further, such a temperature ramping can also damage the nucleic acid of interest and inactivate the enzymes involved in the assay.[14] As a result, specialized thermostable DNA polymerases are also necessary to perform PCR. From a computational perspective, the efficiency of PCR also heavily depends on the primer design process and the optimality of various kinetic parameters of the primer sequence, which are necessary to initiate the extension phase of the PCR cycle. In this context, isothermal NAATs are advantageous over PCR assays as they are operative at a constant temperature, and can amplify the DNA or RNA sequence of interest without considerable denaturation or a mandatory requisite of thermostable polymerases.[12,13] In specific, the RCA method is highly revolutionary in that, it is the first method to consider circular DNA templates as starting material for amplification.[12] The underlying advantages include natural replication of DNA from circular plasmids and viral genomes, possibility of promoter-independent template recognition in the case of RNA polymerases, ability to amplify circular DNA as small as a few tens of nucleotides in size with high efficiency, and combined high specificity and sensitivity with multiplexing ability.[15] These applications demonstrate the potential of RCA as one of the important isothermal NAATs in molecular detection of nucleic acid from a truly diverse pathogen panel of interest. Consequently, RCA has also been extensively utilized to enhance the detection limit of oligonucleotide probes specific to zoonotic pathogens such as newcastle disease virus (NDV),[16] zika virus (ZIKV), ebola virus (EBOV), dengue virus (DENV),[17] etc.

The history of probe design for molecular diagnostics dates back to 1976 when prenatal diagnosis of α-thalassemia was performed based on DNA sequence signatures.[18] With the advent of several advancements in molecular biology, specifically the development of PCR using thermostable DNA polymerase enzymes, the diagnostics industry has seen an explosion of DNA and RNA sequence-based pathogen detection strategies, which were also bolstered further by the completion of the Human Genome Project.[19,20] Today clinical pathology studies in laboratories are heavily reliant on the availability of molecular detection methods specific to pathogens of interest, to arrive at conclusions about disease severity from clinical samples of patients.[20] This is also observable in the COVID-19 pandemic where RT-PCR has become the standard diagnostic test for highly sensitive and specific detection of several viral variants.[21−23] Rapid adaptation of the test procedure to address emerging endemic variants and new waves of the pandemic at a global scale was also observed. All of these demand stringent probe design strategies to further a specific and robust detection of the target of interest. However, with the limitations of PCR-based detection methods outlined earlier and the apparent difficulties in designing new sets of probes for every emerging pathogen and its variants, there is a dire need to automate molecular probe design strategies, which can leverage the potential of RCA in accelerating pathogen detection, such as padlock probes (PLP)[24] or molecular inversion probes (MIP).[25] PLPs add to the levels of specificity in RCA, as it exclusively resorts to circle hybridization.[15] Together, PLP-RCA being an underexplored NAAT method for PoC applications can provide an unprecedented degree of multiplexing, specificity, versatility, and amenability to integration in miniaturized platforms.[15] Herein, we have developed a novel PLP design pipeline called AutoPLP, to automate the design of PLPs for molecular detection of zoonotic pathogens, as a case study.

A plethora of nucleic acid probe design methods have been developed earlier, including Primer3,[26] chipD,[27] OligoMiner,[28] ProbeMaker,[29] PathogenMIPer,[30] and ProbeDealer.[31] However, the existing methods for PLP design can consider only a subset of assay-related technical parameters, speed, time, and memory requirements into consideration, while automating the probe design process. To address this critical need for an advanced and flexible probe design pipeline with general capability to design probes against any given pathogen of interest, AutoPLP is proposed in the current study. AutoPLP improves upon the existing methods by integrating all steps of the PLP design process into a single framework, from sequence data curation to cross-hybridization checks. AutoPLP enables multiplexing at the species level and organism level, by being able to design probes covering user-defined sets of species or serotypes of a pathogen of interest, and for a panel of diverse pathogens, respectively. Unlike the existing PLP design tools such as ProbeMaker[29] and PathogenMIPer,[30] AutoPLP checks for potential secondary structure formation after backbone integration to ensure that the PLPs do not sequester significant stretches of nucleotides from target identification, by self-hybridization. The method has been validated against two common zoonotic pathogens namely, the rabies virus and *Mycobacterium tuberculosis*. Results from AutoPLP show the potential of the method to support and accelerate the development of novel molecular diagnostics against zoonotic pathogens by strategically accounting for all of the relevant experimental parameters already during the design process.

RESULTS AND DISCUSSION

The general features of AutoPLP implementation and the results from both the case studies are discussed in detail in the following sections.

Features of AutoPLP Method. AutoPLP method was implemented in Python (v3.6) and is dependent on BioPython (v1.79), numpy (v1.18.5), ViennaRNA (v2.4.18), and WordCloud (v1.8.1) packages. Further, stand-alone versions of ClustalOmega (v1.2.4), MEGA (v11.0.10), and NCBI BLAST+ (v2.12.0) packages are also required for automated MSA, phylogenetic analysis, and cross-hybridization checkpoints, respectively. Command-line arguments are used to parse the input parameters supplied by the user for each stage, facilitating parallel usage of the AutoPLP method for multiple organisms. The parameters supported by codes corresponding to module 1 and module 3 are tabulated as follows (Tables 1 and 2).

Table 1. Input Parameters Supported by Module 1 of AutoPLP Method with Associated Action Description

flag or input parameter	description	mandatory input
-i (--input)	a file with NCBI accessions to download. If genome size is large, kindly consider using specific gene IDs instead of entire genomes for preprocessing.	yes
-g (--gene_inp)	level of alignment - use "gene" if only a specific gene is used and "genome" if complete genomes are used.	yes
-d (--database)	NCBI database ID (default = nuccore)	no
-e (--email)	an e-mail address for Entrez login	no
-b (--batch)	the number of accessions to process per request (default = 100)	no
-o (--output_dir)	a directory to write downloaded files to	yes
-f (--seqs)	folder with .fasta for each genome or for the gene from related species	no
-s (--sense)	is the genome negative-sense RNA? (yes/no) (default = no)	no
-h (--help)	display all available flags and their description for the user	no

Module 1—pre_process_genomes.py. The automated functionality of module 1 includes sequence data curation, filtering, genome classification, preprocessing, and MSA. Module 1 supports a wide range of input parameters (Table 1) allowing the user to flexibly apply the method depending on the accessibility of sequence data of interest.

Either -i or -f flag, along with -g, must be invoked to use module 1. For negative-sense RNA genomes such as the Ebola virus, module 1 considers both viral RNA and complementary RNA (cRNA) sequences for PLP design and the genome is preprocessed accordingly.

Module 3—gene_level_PLP_design.py. Module 3 is responsible for the extraction and filtration of conserved regions from the MSA, followed by enumeration of all possible backbone combinations, haplotyping, and finalization of PLP sequences for experimental validation. Module 3 takes into account all of the relevant experimental parameters and conditions necessary for the validation of PLPs including T_m, GC content, presence of k-mer repeats, and potential self-hybridization. A list of all of the input parameters supported by module 3 is tabulated below (Table 2).

Module 3 requires a fixed comma-separated variable (CSV) file containing the accession numbers, barcode information, restriction site information, and family assignment obtained from phylogenetic analysis to design PLPs for the target organism. As module 3 is independent of the other modules and vice-versa, sequence data curation and multiple sequence alignment are also possible via module 3, without initially invoking module 1. The output from module 3 is a text file containing the finalized, filtered, consensus PLP sequences for each family of input sequences, along with the appropriate choice of barcode sequences and gene coordinates. The output also includes details on the input parameters utilized for the calculation and the physicochemical properties of the designed PLPs.

Designing Padlock Probes Targeting Rabies Virus. Rabies virus belongs to the *Rhabdoviridae* family and *Lyssavirus* genus. Viruses belonging to the *Lyssavirus* genus are causative agents of viral encephalitis, commonly referred to as Rabies infection. Early detection of rabies in animals and early

Table 2. Input Parameters Supported by Module 3 of AutoPLP Method with Associated Action Description

flag or input parameter	description	mandatory input
-i (--input)	a CSV file with inputs for PLP design (NCBI ID, Barcode1, Restriction site sequence, Barcode2, family)	yes
-s1 (--barcodes)	a file containing barcode sequences of interest	no
-d (--database)	NCBI database ID (default = nuccore)	no
-e (--email)	an e-mail address for Entrez login	no
-b (--batch)	the number of accessions to process per request (default = 100)	no
-o (--output_dir)	length of each PLP arm (in nucleotides) (default = 15). A directory to write downloaded files to	yes
-f (--seqs)	folder with .fasta file(s) or name of a single multi-fasta file with all sequences. If sequences have to be downloaded from NCBI, provide "download" keyword as input	yes
-s2 (--sense)	is the genome negative-sense RNA? (yes/no) (default = no)	no
-l (--arm_length)	length of each PLP arm (in nucleotides) (default = 15)	no
-t (--Tm_thresholds)	lower and upper thresholds for probe T_m (in celcius). Should be given as a pair inside parathesis (t1,t2) where, t1 < t2 (default = (25, 77))	no
-n (--npergene)	number of probes to output for the given set of sequences (default = 3)	no
-c (--conservation)	conservation (in percentage) expected between probe and target. Note that higher percentage can decrease the number of probes output from the code (default = 65%)	no
-g (--gc_content)	lower and upper thresholds for probe G + C content (in percentage). Should be given as a pair inside parathesis (gc1, gc2) where, gc1 < gc2 (default = (30, 50))	no
-k (--kmer_repeats)	length of contiguous single nucleotide repeats (Ex: AAAAAA) allowed in the probe sequence (default = 6)	no
-h (--help)	display all available flags and their description for the user	no

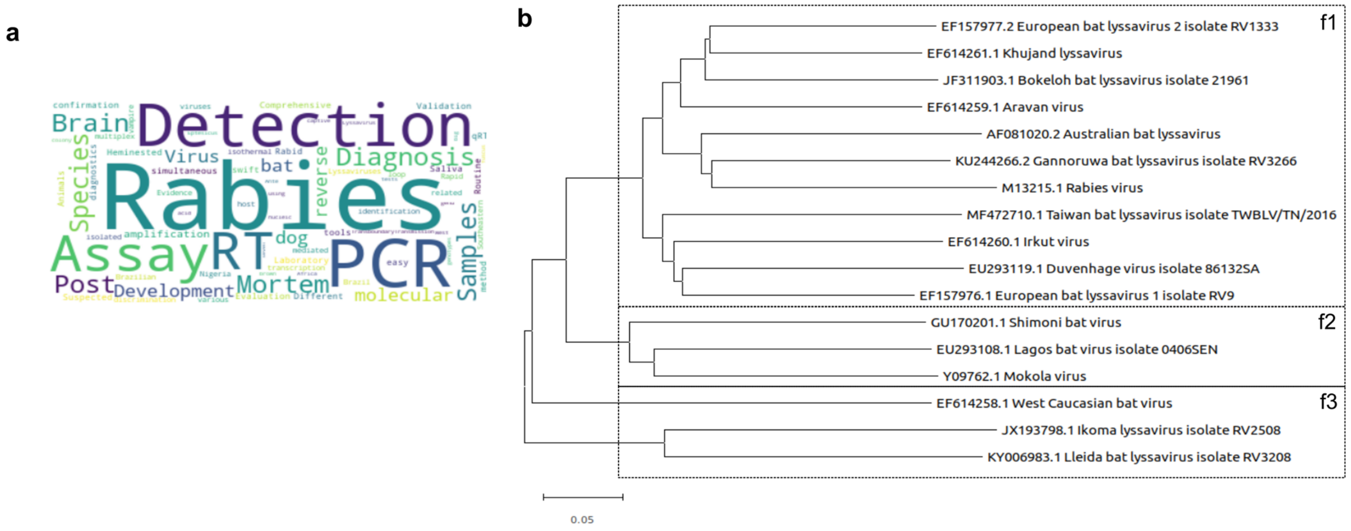

a
confirmation
viruses
Comprehensive
Validation
Brain
Detection
multiplex
qRT
Heminested
Virus
isothermal
Rabid
bat
Diagnosis
simultaneous
swift
Lyssaviruses
Saliva
Rapid
Routine
Species
diagnostics
Animals
Rabies
reverse
Evidence
related
host
identification
isolated
amplification
tools
dog
mediated
Samples
Southeastern
Assay
RT
PCR
Laboratory
transcription
easy
Nigeria
Brazilian
Brazil
method
Post
Development
Mortem
molecular
Suspected
discrimination
various
Evaluation
Different
b
EF157977.2 European bat lyssavirus 2 isolate RV1333
EF614261.1 Khujand lyssavirus
JF311903.1 Bokeloh bat lyssavirus isolate 21961
EF614259.1 Aravan virus
AF081020.2 Australian bat lyssavirus
KU244266.2 Gannoruwa bat lyssavirus isolate RV3266
M13215.1 Rabies virus
MF472710.1 Taiwan bat lyssavirus isolate TWBLV/TN/2016
EF614260.1 Irkut virus
EU293119.1 Duvenhage virus isolate 86132SA
EF157976.1 European bat lyssavirus 1 isolate RV9
GU170201.1 Shimoni bat virus
EU293108.1 Lagos bat virus isolate 0406SEN
Y09762.1 Mokola virus
EF614258.1 West Caucasian bat virus
JX193798.1 Ikoma lyssavirus isolate RV2508
KY006983.1 Lleida bat lyssavirus isolate RV3208
f1
f2
f3
0.05

Figure 1. (a) Word cloud summarizing the information collected on molecular detection of rabies virus from literature survey. (b) Phylogenetic tree obtained from MEGA software using the MSA of rabies genomes generated by AutoPLP Module 1. The three families of rabies virus considered for padlock probe design are designated as f1, f2, and f3 as labeled in the tree. Species of rabies virus belonging to each of the three families are grouped together using dotted boxes.

diagnosis of potential rabies infection in humans can lead to better control of rabies in endemic regions and underdeveloped countries. Several studies showcasing the application of NAATs for the rapid detection of rabies virus from clinical samples have been published. Specifically, the nucleocapsid (N) gene, RNA-dependent RNA polymerase (L) gene, and glycoprotein (G) gene have been extensively studied to extract highly conserved regions, which can serve as targets for sensitive and specific diagnosis of the infection. In this case study, the objective was to design PLP sequences with desired input parameters to accurately detect rabies virus via RCA.

Literature Survey. To extract information on the genes commonly targeted for molecular detection and diagnosis of rabies virus, a detailed literature survey was performed using PubMed keyword search. Genes, probe regions, sensitivity, and specificity of various NAAT methods for the detection of rabies virus were collated. Results from the literature survey were summarized using a word cloud, which records the frequency of various terms commonly observed in titles of articles collated during the literature survey (Figure 1a). Based on the literature survey it was clear that N, L, and G genes in the rabies genome are common target regions for design of primers and probes for molecular detection of rabies infection.

Sequence Data Curation and Preprocessing. Based on the ICTV nomenclature, the NCBI RefSeq accession numbers of the genome sequences of 17 rabies species were obtained (Table 3). With the list of accession numbers, module 1 of AutoPLP was invoked to automatically download the 17 genomes from NCBI Entrez database.[33] The downloaded sequences were preprocessed and subjected to MSA using the ClustalOmega stand-alone program. The resultant MSA output was utilized as input to perform phylogenetic analysis.

Phylogenetic Analysis. The Molecular Evolutionary Genetics Analysis (MEGA) software was used to import the MSA obtained from Module 1 and generate the phylogenetic tree for the set of input sequences. The 17 species were subdivided into three families or taxons based on the clustering observed in the tree (Figure 1b). According to the families assigned to each viral species, an input file was prepared for the design of PLPs using module 3.

Table 3. NCBI Accession Numbers of the Genome Sequences of 17 Species of Lyssaviruses Obtained from the ICTV Nomenclature

NCBI accession number	species
EF614259	Aravan lyssavirus
AF081020	Australian bat lyssavirus
JF311903	Bokeloh bat lyssavirus
EU293119	Duvenhage lyssavirus
EF157976	European bat 1 lyssavirus
EF157977	European bat 2 lyssavirus
KU244266	Gannoruwa bat lyssavirus
JX193798	Ikoma lyssavirus
EF614260	Irkut lyssavirus
EF614261	Khujand lyssavirus
EU293108	Lagos bat lyssavirus
KY006983	Lleida bat lyssavirus
Y09762	Mokola lyssavirus
M13215	Rabies lyssavirus
GU170201	Shimoni bat lyssavirus
MF472710	Taiwan bat lyssavirus
EF614258	West Caucasian bat lyssavirus

Padlock Probe Design and Filtration. As discussed earlier, a set of finalized parameters (Table 5) were used to design PLPs for the case studies. Using the input file prepared based on the results from phylogenetic analysis and a set of backbone sequences and restriction sites of interest (see Supporting Information 2), module 3 of AutoPLP was designed to enumerate all possible combinations of backbone sequences, which are stitched to the probe regions identified based on sequence conservation, GC content, melting temperature (T_m) and other filtration parameters. This resulted in a final set of 207, 216, and 29 probes for N, L, and G genes, respectively. It must be noted that the probes output by module 3 are consensus probe sequences, composed of nonstandard nucleotides, haplotyping (see Nomenclature Appendix) the variation or degeneracy of nucleotides in each position of the

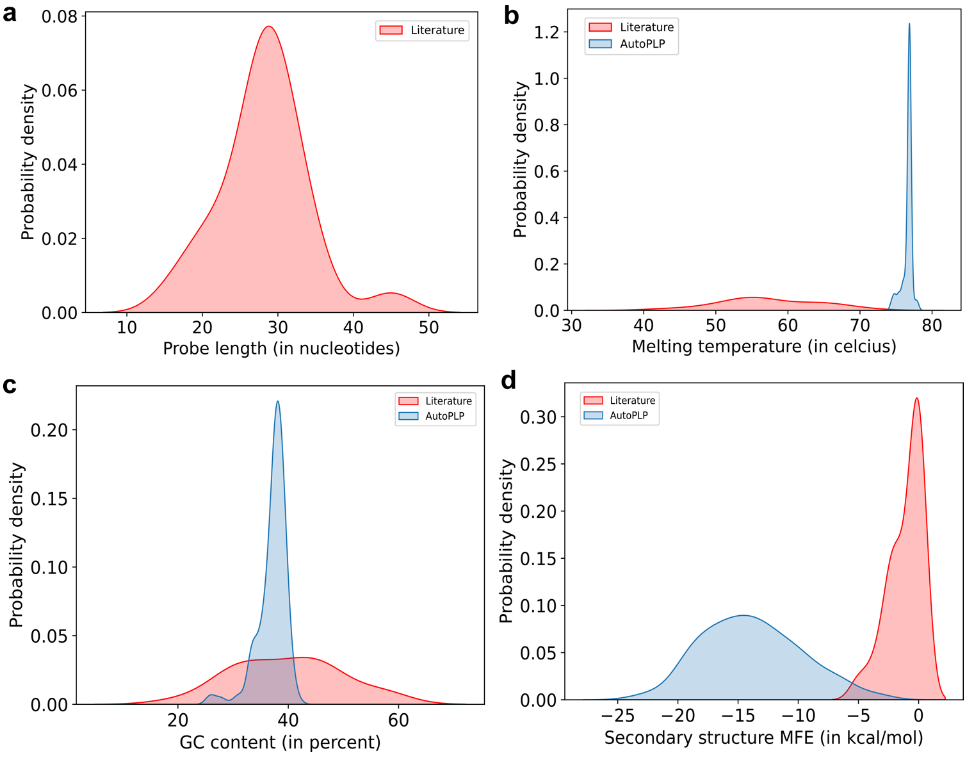


Figure 2. Kernel density estimation (KDE) plots showing the distribution of various key properties of the probe sequences collected from the literature (red) in comparison with that of the PLPs designed by AutoPLP (blue) for the rabies virus: (a) probe length; (b) T_m (in celcius); (c) GC content (in %); (d) MFE (in kcal/mol).

probe. The final set of consensus PLPs obtained from AutoPLP for rabies virus are provided as part of the Supporting Information 2. The formats followed for input and output files in AutoPLP method are also outlined as part of the Supporting information 1 (Section S1).

Analysis of Filtered Probe Sequences. The final set of probe sequences obtained from AutoPLP method for each gene of interest in the rabies virus was compared with analogous probe sequences for molecular detection of rabies virus, obtained from literature survey, thereby providing a second level of confirmation for the target binding sequence information. A significant difference in the size of probe sequences was observed in the literature, as most of the probes were designed specifically to RT-PCR assay and were quite smaller compared to the size of the PLPs from AutoPLP (Figure 2a). It is notable that the PLPs from AutoPLP had a constant length of 93 nucleotides. Further, key properties of the designed probes including T_m, GC content, and minimum free energy (MFE) of secondary structure formation were compared with that of the probes collected from the literature (Figure 2b−d).

The T_m distribution (Figure 2b) indicates that the PLPs designed by AutoPLP have a high and narrow range of T_m compared to the probes collected from the literature. This can be attributed to the wide difference in the length of probe sequences being compared, which is known to impact the T_m of the sequence due to a potential increase in hydrogen bonding with the increase in sequence length—for instance, observable differences between T_m for binding regions only versus T_m for the full PLP length (see Supporting information 1—Figure S3).[32] The GC content (Figure 2c) of both datasets of probes was found to lie within 20−60% as desirable. Similar argument can also be made for the observed difference in MFE distributions (Figure 2d) of the two sets of probe sequences.[33] The observed differences in Figure 2D can be attributed to the relatively lower number of sequences available in the literature for rabies virus being analyzed. However, when extended to probe sets with increased number of datasets, the differences can be more pronounced with a noticeable shift in the MFE levels toward the lower end (see Supporting information 1—Figure S4).

Designing Padlock Probes Targeting *M. Tuberculosis*.

Mycobacterium tuberculosis (*Mtb*) is a pathogen of major concern due to its ability to withstand both first- and second-line antibiotics, leading to drug resistance. It can also affect multiple organ systems of the human body, including spine, kidney, and brain, leading to extra-pulmonary tuberculosis (EPTB).[34] Clinically, two types of tuberculosis have been identified based on the advent of the infection after exposure to the pathogen: latent tuberculosis infection (LTBI) and tuberculosis disease (TB). In the case of LTBI, the infection is asymptomatic despite exposure to the pathogen, making diagnosis and treatment equally challenging.[35] As of 2020, an estimated 5.8 million people have been infected with TB globally, with India, Indonesia, and the Philippines being the worst affected countries.[36] The primary genes of interest to detect emerging extensively drug-resistant TB strains (XDR-TB) are catalase peroxidase (*katG*) and DNA-dependent RNA polymerase β subunit (*rpoB*) which lead to resistance for the TB drugs Isoniazid and Rifampicin, respectively. Other target genes commonly found in the literature for molecular detection of *M. tuberculosis* are *rrs* (16S rRNA gene), *devR*, *fabG1*, *gyrA*, *gyrB*, *inhA*, insertion sequences (IS1081, IS6110, IS986), *mpb64*, *rimM*, and *sdaA*. In this case study, the objective was to design PLP sequences with desired input

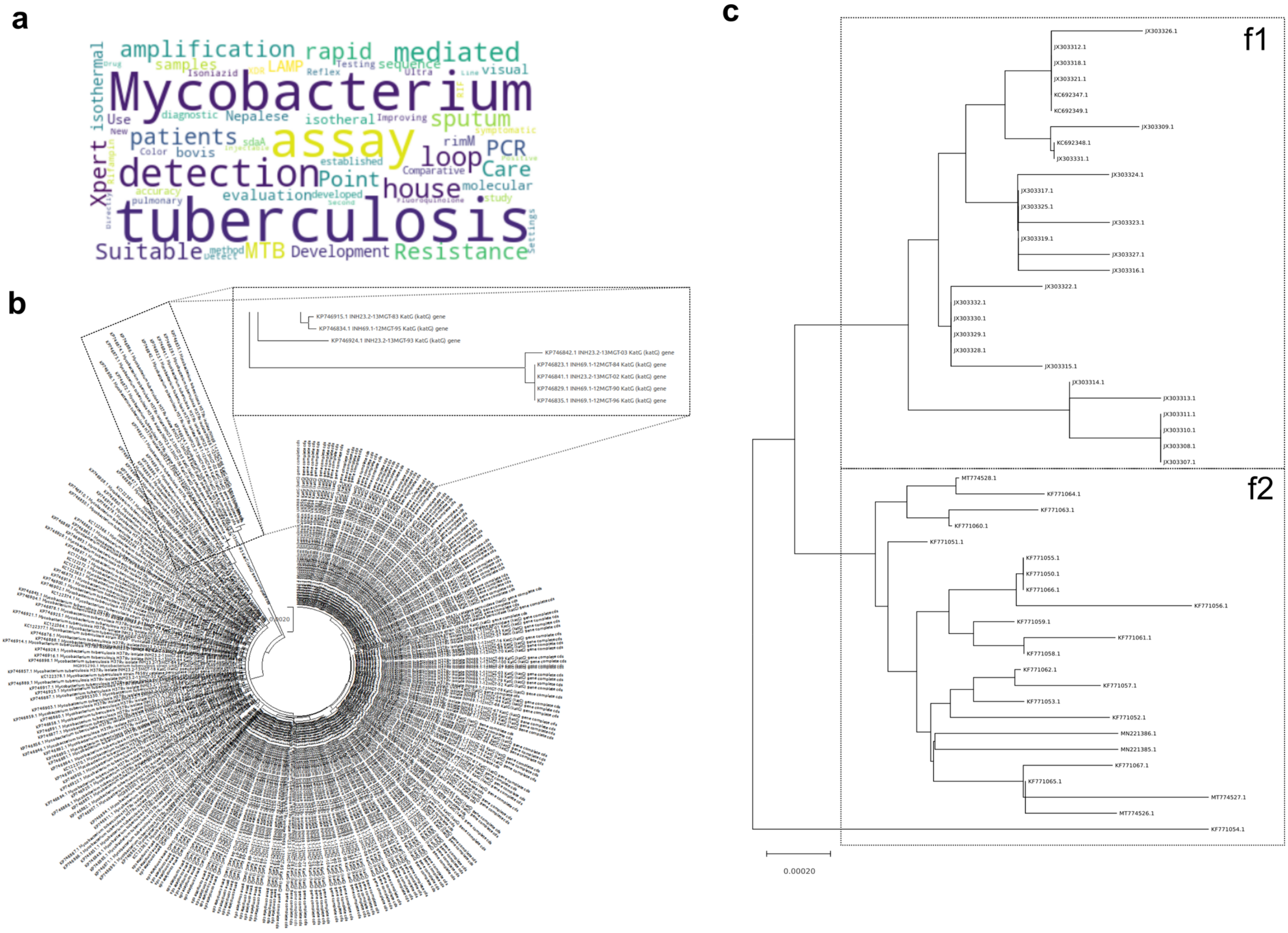


Figure 3. (a) Word cloud summarizing the information collected on molecular detection of *M. tuberculosis* from literature survey. Phylogenetic tree obtained from MEGA software using the MSA of (b) *katG* and (c) *rpoB* gene sequences generated by AutoPLP Module 1. Due to the dense clustering observed in the *katG* phylogenetic tree, the family definitions considered for probe design are not shown in the figure. A small inset shows one of the prominent sub-branches in the tree. The two families of *rpoB* sequences considered for padlock probe design are designated as f1 and f2 as labeled in the tree.

parameters to accurately detect drug-resistant TB strains via RCA, by targeting the *katG* and *rpoB* genes.

Literature Survey. To extract information on the genes commonly targeted for molecular detection and diagnosis of *Mtb*, a detailed literature survey was performed using PubMed keyword search. Further, TB-specific databases such as the TB Database (TBDB)[37] and TB portals[38] were used to obtain information on various drug-resistant TB strains of interest. Genes, probe regions, sensitivity, and specificity of various NAAT methods for the detection of TB were collated. Results from the literature survey were summarized using a word cloud (Figure 3a).

Sequence Data Curation and Preprocessing. To target the *katG* and *rpoB* genes of drug-resistant TB strains, PopSet datasets of *katG* and *rpoB* sequences were collected from the NCBI website and converted into a list of accession numbers for download. This amounted to a set of 278 *katG* and 51 *rpoB* sequences (Table 4). With the list of accession numbers (Supporting information 1, Table S2), module 1 of AutoPLP was invoked to automatically download the gene sequences from NCBI Entrez database.[42] The downloaded sequences were preprocessed and subjected to MSA using the ClustalOmega stand-alone program. The resultant MSA output was utilized as input to perform phylogenetic analysis.

Table 4. NCBI PopSet Dataset Identifiers of the *katG* and *rpoB* Gene Sequences from Drug-Resistant *M. tuberculosis* Strains Utilized for the Case Study

gene	NCBI PopSet ID	no. of sequences
katG	449838665	16
	929524375	187
	1373737810	75
rpoB	405113443	25
	478718539	3
	569533973	18
	1905476421	3
	1917459101	2

Phylogenetic Analysis. The Molecular Evolutionary Genetics Analysis (MEGA) software was used to import the MSA obtained from Module 1 and generate the phylogenetic tree for the set of input sequences. The neighbor-joining method with Tamura−Nei distance model was used to build the phylogenetic tree, and the *katG* and *rpoB* gene sequences were subdivided into three and two families, respectively, based on the clustering observed in the tree (Figure 3b,c).

According to the families assigned to each gene sequence, an input file was prepared for the design of PLPs using module 3.

Padlock Probe Design and Filtration. As discussed earlier, a set of finalized parameters (Table 5) were used to design

Table 5. Input Parameters for the AutoPLP Method for Design of PLPs for the Case Studies Showcased in This Work[a]

parameter	rabies virus	*M. tuberculosis*
negative-sense RNA genome	no	no
length of PLP arms (in nt)	20	20
T_m thresholds (in °C)	(25, 77)	(25, 85)
minimum number of probes to output per gene	3	3
sequence conservation threshold (in %)	65	50
GC content threshold (in %)	(30, 50)	(30, 50)
repeated bases threshold (in nt)	6	6

[a] nt—nucleotides.

PLPs for the case studies. Using the input file prepared based on the results from phylogenetic analysis and a set of backbone sequences encompassing detection and restriction sites of interest, module 3 of AutoPLP was designed to enumerate all possible combinations of backbone sequences, which are stitched to the probe regions identified based on sequence conservation, GC content, melting temperature (T_m) and other filtration parameters. This resulted in a final set of 11 and 2 probes for *katG* and *rpoB* genes, respectively. It must be noted that the probes output by module 3 are consensus probe sequences, composed of nonstandard nucleotides haplotyping (see Nomenclature Appendix) the variation or degeneracy of nucleotides in each position of the probe. The final set of consensus PLPs obtained from AutoPLP for *Mtb* are provided as part of the Supporting Information 2. The formats followed for input and output files in AutoPLP method are also outlined as part of the Supporting information 1 (Section S1).

Future Perspectives. The AutoPLP pipeline currently provides automated modules for over 70% of the functionalities necessary for successful PLP design experiments, reducing the manual intervention required in the process. However, the key element of PLP design, namely, the target region or gene selection, remains in the manual module (Module 2) to facilitate flexibility to the user in considering tested target regions from an extensive literature survey. The upcoming version(s) of AutoPLP will be aimed at automating the literature mining analysis with machine learning and artificial intelligence-based methods,[39] to automate the target region selection for PLP design. The identified target regions will be at user discretion within Module 2. Further, some advanced parameters such as the sinusoidal template length-dependent amplification bias observed in RCA,[40] will be considered in Module 3, to suggest optimal PLP length for amplification of longer probe sequences.

■ CONCLUSIONS

In this study, a novel method was developed to enable rapid design of PLPs for diagnosis of zoonotic infections. The method automates key phases of the probe design pipeline in three modules and provides control over multiple experimental parameters of interest, which can increase the *in vitro* detection efficiency of the designed PLP. *In silico* validation of the method was performed by designing PLPs targeting multiple genes of interest in Rabies virus and *M. tuberculosis*. Specifically, probes were designed against genes responsible for antimicrobial resistance in multiple strains of *M. tuberculosis*, to test the ability of the method when applied on sequences of high diversity. The filtered set of probes from the method were further compared with existing probes collected

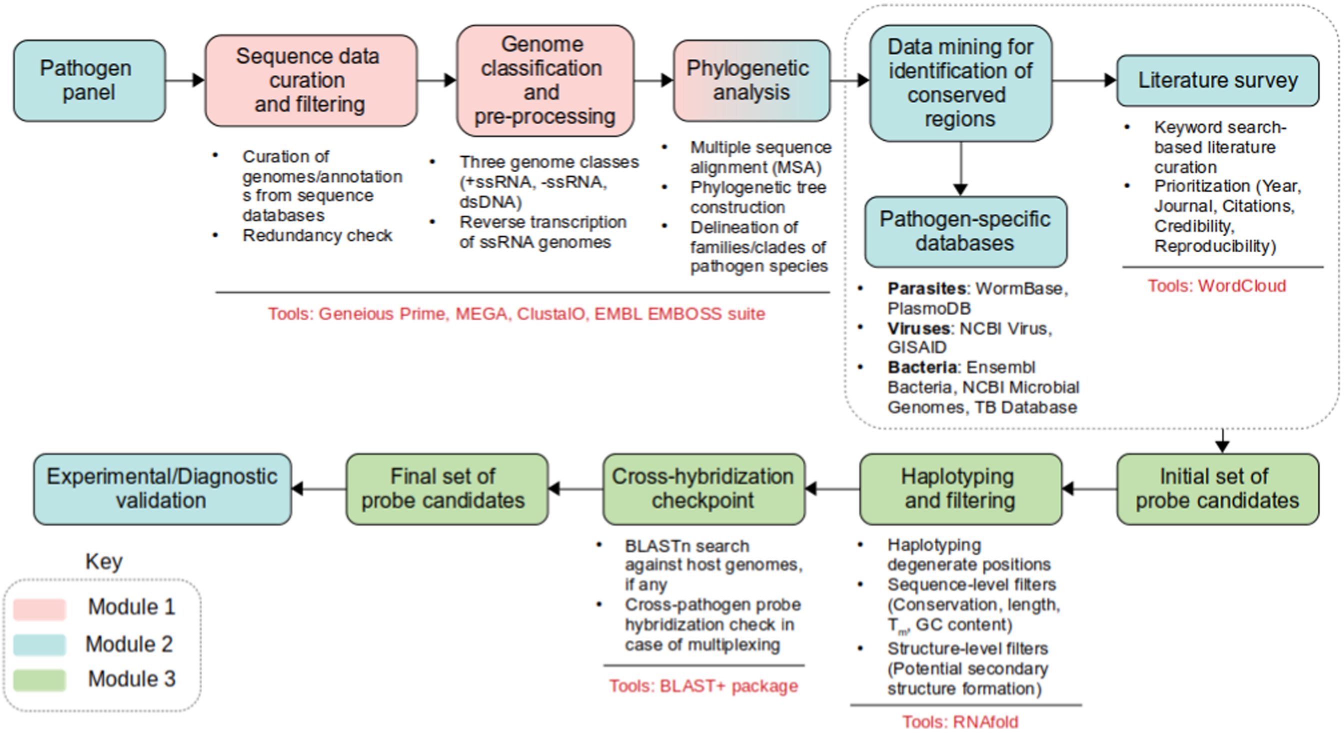

Pathogen panel
Sequence data curation and filtering
Genome classification and pre-processing
Phylogenetic analysis
Data mining for identification of conserved regions
Literature survey
• Curation of genomes/annotations from sequence databases
• Redundancy check
• Three genome classes (+ssRNA, -ssRNA, dsDNA)
• Reverse transcription of ssRNA genomes
• Multiple sequence alignment (MSA)
• Phylogenetic tree construction
• Delineation of families/clades of pathogen species
Tools: Geneious Prime, MEGA, ClustalO, EMBL EMBOSS suite
Pathogen-specific databases
• Parasites: WormBase, PlasmoDB
• Viruses: NCBI Virus, GISAID
• Bacteria: Ensembl Bacteria, NCBI Microbial Genomes, TB Database
• Keyword search-based literature curation
• Prioritization (Year, Journal, Citations, Credibility, Reproducibility)
Tools: WordCloud
Experimental/Diagnostic validation
Final set of probe candidates
Cross-hybridization checkpoint
Haplotyping and filtering
Initial set of probe candidates
• BLASTn search against host genomes, if any
• Cross-pathogen probe hybridization check in case of multiplexing
Tools: BLAST+ package
• Haplotyping degenerate positions
• Sequence-level filters (Conservation, length, Tm, GC content)
• Structure-level filters (Potential secondary structure formation)
Tools: RNAfold
Key
Module 1
Module 2
Module 3

Figure 4. Overall architecture of AutoPLP with the components of each primary module shown in different colors.

from the literature in terms of their property profiles. The results show the potential of this method to accelerate the development of novel molecular diagnostics against zoonotic pathogens by accounting for all of the relevant experimental parameters during the design process.

■ MATERIALS AND METHODS

AutoPLP is a novel pipeline developed in this study to facilitate rapid design of PLPs or MIPs for the detection of any pathogen of interest (for details on the interaction of PLPs with target genes of interest, see refs 15, 16). The pipeline has been developed in a modular manner with three primary modules as described below. *In silico* validation of AutoPLP was performed through the design of PLPs for a hypervariable virus and a multidrug-resistant bacterium, as part of a zoonotic pathogen panel.

Overview of AutoPLP. The overall architecture of AutoPLP consists of several components, which can be grouped under three primary modules (Figure 4).

Module 1. The objective of module 1 is to collect sequence information (genomes or gene sequences) for various species of a pathogen of interest and establish the taxonomical hierarchy of the pathogen through phylogenetic analysis. To achieve this, module 1 utilizes the following three sub-modules.

- ***Sequence data curation and filtering***: This sub-module takes the NCBI accession identifiers for the genomes or genes of interest as input, for a list of species/strains for every pathogen of a panel. The genome or gene sequences are automatically downloaded in FASTA format,[41] into a user-defined path in the system by connecting with the NCBI Entrez database[42] using BioPython.[43] The user can also entirely bypass this step by providing a path to a folder containing the genome or gene sequences in FASTA format. This is specifically useful in cases where the sequence information has to be curated from alternate sources other than the NCBI Entrez database.[42] An initial filtering is done at this point by a redundancy check for the removal of any duplicated sequence information.
- ***Genome classification and preprocessing***: This sub-module classifies the input genome sequences into one of the following three categories: negative-sense single-stranded RNA genomes, positive-sense single-stranded RNA genomes, and double-stranded genomes. This step is essential in the case of pathogen panels involving viruses, where negative-sense RNA genomes are possible, such as the Ebola virus.[17] For positive-sense single-stranded RNA genomes, the genome sequence can be directly used in probe design. For negative-sense single-stranded RNA genomes, the genome is considered as viral RNA (vRNA), with polarity opposite to that of the viral mRNA or complementary RNA (cRNA). Therefore, the module performs reverse transcription on the vRNA to obtain the cRNA, with the same polarity as that of the viral mRNA. During the later stages of PLP design, probes will be designed against both vRNA and cRNA sequences for organisms with negative-sense single-stranded RNA genome.
- ***Multiple sequence alignment (MSA)***: In this penultimate sub-module of module 1, an MSA is performed for the curated genome or gene sequences using the ClustalOmega program[44] from the EMBL EMBOSS suite.[45] The gap open and gap extension penalties for the MSA are kept at their default values, but can be modified as per user preference. The FASTA format output from ClustalOmega can be used as input to Module 2 of AutoPLP, which is described below. While using module 1, it is advisable to supply genome sequences for performing the MSA only if the genome size is less than 20 kilo-base pairs (kbp). Otherwise, supplying the sequences of a gene of interest for which PLPs have to be designed, rather than the whole genome sequence, will provide a better trade-off between the time taken for MSA and the PLP design. It is notable that this choice is at the discretion of the user depending on the pathogen, time, and computational resources available to design PLPs.

Module 2. This is the only module requiring manual intervention in AutoPLP. The objective of this module is twofold: delineating the taxonomical hierarchy of the pathogen using the MSA output from module 1, and mining existing literature and pathogen-specific databases for identification of highly conserved regions in their genes or genomes, which are potential target regions for PLP design. Likewise, for emerging and reemerging strains of pathogens, the hypervariable regions responsible for the generation of novel variants can also be effectively targeted using this approach.

- ***Phylogenetic analysis***: The MSA output in FASTA format obtained from module 1 can be used as input to existing phylogenetic analysis tools such as MEGA[46] or the commercially available Geneious Prime. With a suitable choice of evolutionary distance model and tree-building method available in the tool, the phylogenetic tree can be obtained for the set of gene or genome sequences considered as input. For the validation of AutoPLP, MEGA was used as the tool for phylogenetic analysis with the Tamura–Nei distance model,[47] to construct dendrograms using the neighbor-joining method. Using standard tree-cut approaches available in the literature to define groups of taxa, the different sub-families present in the organism can be identified.
- ***Data mining for identification of conserved regions***: One of the critical steps of PLP design for the detection of a pathogen is the identification of the target gene or genetic region of interest. Selection of such regions can be made by identification of pathogen-associated molecular patterns (PAMPs), which are conserved regions of the pathogen genome coding for gene products capable of triggering the innate immune response in the host immune system.[48] Based on existing literature on NAAT for oligonucleotide-based pathogen detection, such highly conserved regions within the pathogen genome or pathogen-specific genes can be identified, which are ideal candidates for probe design. Such genes or genetic regions should not have any considerable similarity to the host genome (human genome in most cases), to ensure specificity of the PLPs in binding to the target region of the pathogen. Apart from literature mining, conserved genes and genetic regions can also be identified using pathogen-specific databases. The results from this sub-module can be visualized using word clouds to understand the relative frequency of observation of various pathogen detection methods.

Module 3. The objective of this module is to filter, freeze, and finalize (FFF) the PLP sequences for a target gene of interest in a pathogen.

- ***Initial set of probe candidates***: Based on the taxonomical hierarchy for the pathogenic species/strains obtained from module 1 and a target gene identified from module 2, this sub-module can extract subsequences of user-defined length from the target gene sequences for each sub-family of the organism. This is achieved using a sliding-window approach where the window size is controlled based on the user-defined parameter "*Arm length*", referring to the length of oligonucleotide bases involved in the target binding.
- ***Haplotyping and filtering***: Haplotyping (see Nomenclature Appendix) is used to cluster groups of degenerate target sequences into a consensus sequence, against which the probes will be designed. Positions of variation in the consensus sequence can be substituted with Wobble base pairs to enhance the variation tolerance of the designed probes. To further facilitate the direct applicability of the extracted probe sequences for the development of an isothermal NAAT, several filtration parameters are applied on both the probe and target sequences, which are described below.
    - ***Probe length***: Length of the 5′ and 3′ arms of the PLP. Default value is set to 15 nucleotides per arm, capable of hybridizing to a target region of 30 nucleotides in length.
    - ***Sequence conservation***: Percentage conservation of nucleotide positions expected in the target sub-sequence. The default conservation percentage is set to 65% of the target sub-sequence length, as the probability of observing a longer conserved region in the target gene can decrease drastically as the number of sequences or species in a sub-family of the organism increases, due to alignment errors leading to ambiguously aligned regions (AARs).[49,50]
    - ***T_m threshold***: Melting temperature (T_m) of the probe sequence, initially done for the target binding region, and then extended to the entire PLP length including the backbone. Default range for T_m is set from 25 to 77 °C, to allow a wide operating temperature for the probe sequence without denaturation from the hybridized state. In specific, the lower range caters to experimental conditions involving room temperature which is more relevant for isothermal NAATs, while the upper range is based on the operating temperature of the enzymes used in experiments. A more stringent threshold can be enforced during post-processing of the finalized probe sequences, to obtain a smaller subset of probes with high confidence. T_m is calculated based on an implementation of the nearest-neighbor thermodynamics method.
    - ***GC content***: Although T_m and GC content are highly correlated, this filtration parameter is explicitly enforced to increase the stringency of the output probes from AutoPLP. The default range of percentage GC content of the target sequence is set from 30 to 50% and can be modified as per user requirement to tailor the pipeline for the organism of interest.
    - ***Self-hybridization potential***: Percentage of nucleotides in the target sub-sequence and probe sequence, which can be tolerated in a potential secondary structure formed by base-pairing with itself (self-hybridization). The default value is set to a maximum of 30% of nucleotides in the stem region of the predicted secondary structure for the target sequence or the probe sequence. The entire PLP length has been considered in the current version for the internal loop formation check; however, as required for mutation studies, emphasis can be laid specifically on the 3′ or 5′ ends and the associated loop formation therein. The RNAfold algorithm in the ViennaRNA python package is used to predict the secondary structure for a given oligonucleotide sequence.
- ***Cross-hybridization checkpoint***: The probes obtained after filtration must be checked for potential cross-hybridization with the host genome (in most cases, human genome), to prevent the probes from generating false positive or spurious signals during pathogen detection *in vitro* by experimental validation. In the event of multiplexing, it is also necessary to check for cross-hybridization of filtered probes from each organism with the genomes of other organisms both within and outside the panel considered for study. This will ensure that the probes are specific enough to identify only the pathogen of interest, and will not generate false positives due to cross-pathogen targeting.

The final set of probe candidates obtained from AutoPLP will be directly amenable to experimental validation for highly specific and sensitive detection of the pathogen of interest from clinical samples. Apart from the experimental parameters considered in the current version of AutoPLP, it would also be interesting to consider advanced criteria such as ligase-specific footprints to allow for potential regions of mismatches between the PLP and the target region, and application of a composite scoring function to rank the output PLPs for further experimental validation. Future versions of AutoPLP can also include a specific analysis sub-module to compare the critical parameters of the designed PLPs with that of the nucleic acid probes existing in the literature for the pathogenic gene of interest.

Case Studies Using AutoPLP. To showcase the application of the AutoPLP method, a panel of two pathogenic micro-organisms, namely, the rabies virus and *M. tuberculosis* are considered for PLP design. The following parameter settings are used for probe design using AutoPLP for each organism (Table 5). The changes in parameter values between the two organisms are made to accommodate the differences observed in sequence diversity.

The resultant probes are compared with the existing oligonucleotide probes in literature, to elucidate the differences in the parameter distributions of the designed PLPs. Further, the time taken by AutoPLP to design PLPs for the two organisms is also documented.

ASSOCIATED CONTENT

Data Availability Statement

The source code for AutoPLP method will be made available upon request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsinfecdis.2c00436.

> **Supporting information 1**: Input and output file formats supported by AutoPLP, list of noncanonical alphabets used in nucleotide nomenclature, list of accession numbers of *Mtb katG* and *rpoB* sequences used in the case study, Nomenclature Appendix, along with supporting figures (PDF)
>
> **Supporting information 2**: List of chosen backbone sequences and final list of padlock probe sequences designed by the AutoPLP method for target genes of interest in Rabies virus and *Mtb* (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Narayanan Madaboosi − *Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India*; Email: narayananms@iitm.ac.in

M. Michael Gromiha − *Protein Bioinformatics Lab, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India; International Research Frontiers Initiative, School of Computing, Tokyo Institute of Technology, Yokohama 226-8501, Japan*; orcid.org/0000-0002-1776-4096; Email: gromiha@iitm.ac.in

Authors

Sowmya Ramaswamy Krishnan − *Protein Bioinformatics Lab, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India; TCS Research (Life Sciences Division), Tata Consultancy Services, Hyderabad 500081, India*; orcid.org/0000-0001-5404-3266

Ruben R. G. Soares − *Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Solna SE-17121, Sweden*; orcid.org/0000-0001-5958-5232

Complete contact information is available at: https://pubs.acs.org/10.1021/acsinfecdis.2c00436

Notes

The authors declare no competing financial interest.
S.R.K. is working as a researcher at Tata Consultancy Services Limited.

ACKNOWLEDGMENTS

The authors thank the Bioinformatics Infrastructure facility, Department of Biotechnology, and Indian Institute of Technology Madras for computational facilities. N.M. acknowledges support from the Department of Biotechnology, Govt. of India, for Ramalingaswami Re-Entry Fellowship 2020−2021.

REFERENCES

(1) Hassell, J. M.; Begon, M.; Ward, M. J.; Fèvre, E. M. Urbanization and Disease Emergence: Dynamics at the Wildlife-Livestock-Human Interface. *Trends Ecol. Evol.* **2017**, *32*, 55−67.

(2) Hogan, A. B.; Jewell, B. L.; Sherrard-Smith, E.; Vesga, J. F.; Watson, O. J.; Whittaker, C.; Hamlet, A.; Smith, J. A.; Winskill, P.; Verity, R.; Baguelin, M.; Lees, J. A.; Whittles, L. K.; Ainslie, K. E. C.; Bhatt, S.; Boonyasiri, A.; Brazeau, N. F.; Cattarino, L.; Cooper, L. V.; Coupland, H.; Cuomo-Dannenburg, G.; Dighe, A.; Djaafara, B. A.; Donnelly, C. A.; Eaton, J. W.; van Elsland, S. L.; FitzJohn, R. G.; Fu, H.; Gaythorpe, K. A. M.; Green, W.; Haw, D. J.; Hayes, S.; Hinsley, W.; Imai, N.; Laydon, D. J.; Mangal, T. D.; Mellan, T. A.; Mishra, S.; Nedjati-Gilani, G.; Parag, K. V.; Thompson, H. A.; Unwin, H. J. T.; Vollmer, M. A. C.; Walters, C. E.; Wang, H.; Wang, Y.; Xi, X.; Ferguson, N. M.; Okell, L. C.; Churcher, T. S.; Arinaminpathy, N.; Ghani, A. C.; Walker, P. G. T.; Hallett, T. B. Potential impact of the COVID-19 pandemic on HIV, tuberculosis, and malaria in low-income and middle-income countries: a modelling study. *Lancet Global Health* **2020**, *8*, e1132−e1141.

(3) Morens, D. M.; Fauci, A. S. Emerging Pandemic Diseases: How We Got to COVID-19. *Cell* **2020**, *182*, 1077−1092.

(4) Larsson, D. G. J.; Flach, C. Antibiotic resistance in the environment. *Nat. Rev. Microbiol.* **2022**, *20*, 257−269.

(5) Minot, S.; Grunberg, S.; Wu, G. D.; Lewis, J. D.; Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 3962−3966.

(6) Muñoz-Chimeno, M.; Cenalmor, A.; Garcia-Lugo, M. A.; Hernandez, M.; Rodriguez-Lazaro, D.; Avellon, A. Proline-Rich Hypervariable Region of Hepatitis E Virus: Arranging the Disorder. *Microorganisms* **2020**, *8*, 1417.

(7) Carrasco-Hernandez, R.; Jácome, R.; Vidal, Y. L.; de León, S. P. Are RNA Viruses Candidate Agents for the Next Global Pandemic? A Review. *ILAR J.* **2017**, *58*, 343−358.

(8) Bird, B. H.; Mazet, J. A. K. Detection of Emerging Zoonotic Pathogens: An Integrated One Health Approach. *Annu. Rev. Anim. Biosci.* **2018**, *6*, 121−139.

(9) Kang, T.; Lu, J.; Yu, T.; Long, Y.; Liu, G. Advances in nucleic acid amplification techniques (NAATs): COVID-19 point-of-care diagnostics as an example. *Biosens. Bioelectron.* **2022**, *206*, No. 114109.

(10) Ferone, M.; Gowen, A.; Fanning, S.; Scannell, A. G. M. Microbial detection and identification methods: Bench top assays to omics approaches. *Compr. Rev. Food Sci. Food Saf.* **2020**, *19*, 3106−3129.

(11) Vainionpää, R.; Waris, M.; Leinikki, P. Diagnostic Techniques: Serological and Molecular Approaches *Reference Module in Biomedical Sciences* 2015, DOI: 10.1016/B978-0-12-801238-3.02558-7.

(12) Mohsen, M. G.; Kool, E. T. The Discovery of Rolling Circle Amplification and Rolling Circle Transcription. *Acc. Chem. Res.* **2016**, *49*, 2540−2550.

(13) Obande, G. A.; Singh, K. K. B. Current and Future Perspectives on Isothermal Nucleic Acid Amplification Technologies for Diagnosing Infections. *Infect. Drug Resist.* **2020**, *13*, 455−483.

(14) Kantidze, O. L.; Velichko, A. K.; Luzhin, A. V.; Razin, S. V. Heat Stress-Induced DNA Damage. *Acta Nat.* **2016**, *8*, 75−78.

(15) Soares, R. R. G.; Madaboosi, N.; Nilsson, M. Rolling Circle Amplification in Integrated Microsystems: An Uncut Gem toward Massively Multiplexed Pathogen Diagnostics and Genotyping. *Acc. Chem. Res.* **2021**, *54*, 3979−3990.

(16) Ciftci, S.; Neumann, F.; Hernández-Neuta, I.; Hakhverdyan, M.; Bálint, Á.; Herthnek, D.; Madaboosi, N.; Nilsson, M. A novel mutation tolerant padlock probe design for multiplexed detection of hypervariable RNA viruses. *Sci. Rep.* **2019**, *9*, No. 2872.

(17) Ciftci, S.; Neumann, F.; Abdurahman, S.; Appelberg, K. S.; Mirazimi, A.; Nilsson, M.; Madaboosi, N. Digital Rolling Circle Amplification-Based Detection of Ebola and Other Tropical Viruses. *J. Mol. Diagn.* **2020**, *22*, 272−283.

(18) Kan, Y. W.; Dozy, A. M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5631−5635.

(19) Procop, G. W. Molecular diagnostics for the detection and characterization of microbial pathogens. *Clin. Infect. Dis.* **2007**, *45*, S99−S111.

(20) Patrinos, G. P.; Danielson, P. B.; Ansorge, W. J. Chapter 1 - Molecular Diagnostics: Past, Present, and Future. *Mol. Diagn.* **2017**, 1−11.

(21) Feng, W.; Newbigging, A. M.; Le, C.; Pang, B.; Peng, H.; Cao, Y.; Wu, J.; Abbas, G.; Song, J.; Wang, D.; Cui, M.; Tao, J.; Tyrrell, D. L.; Zhang, X.; Zhang, H.; Le, X. C. Molecular Diagnosis of COVID-19: Challenges and Research Needs. *Anal. Chem.* **2020**, *92*, 10196−10209.

(22) Carter, L. J.; Garner, L. V.; Smoot, J. W.; Li, Y.; Zhou, Q.; Saveson, C. J.; Sasso, J. M.; Gregg, A. C.; Soares, D. J.; Beskid, T. R.; Jervey, S. R.; Liu, C. Assay Techniques and Test Development for COVID-19 Diagnosis. *ACS Cent. Sci.* **2020**, *6*, 591−605.

(23) Udugama, B.; Kadhiresan, P.; Kozlowski, H. N.; Malekjahani, A.; Osborne, M.; Li, V. Y. C.; Chen, H.; Mubareka, S.; Gubbay, J. B.; Chan, W. C. W. Diagnosing COVID-19: The Disease and Tools for Detection. *ACS Nano* **2020**, *14*, 3822−3835.

(24) Nilsson, M.; Malmgren, H.; Samiotaki, M.; Kwiatkowski, M.; Chowdhary, B. P.; Landegren, U. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **1994**, *265*, 2085−2088.

(25) Hardenbol, P.; Banér, J.; Jain, M.; Nilsson, M.; Namsaraev, E. A.; Karlin-Neumann, G. A.; Fakhrai-Rad, H.; Ronaghi, M.; Willis, T. D.; Landegren, U.; Davis, R. W. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **2003**, *21*, 673−678.

(26) Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B. C.; Remm, M.; Rozen, S. G. Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **2012**, *40*, No. e115.

(27) Dufour, Y. S.; Wesenberg, G. E.; Tritt, A. J.; Glasner, J. D.; Perna, N. T.; Mitchell, J. C.; Donohue, T. J. chipD: a web tool to design oligonucleotide probes for high-density tiling arrays. *Nucleic Acids Res.* **2010**, *38*, W321−5.

(28) Beliveau, B. J.; Kishi, J. Y.; Nir, G.; Sasaki, H. M.; Saka, S. K.; Nguyen, S. C.; Wu, C.; Yin, P. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E2183−E2192.

(29) Stenberg, J.; Nilsson, M.; Landegren, U. ProbeMaker: an extensible framework for design of sets of oligonucleotide probes. *BMC Bioinformatics* **2005**, *6*, No. 229.

(30) Thiyagarajan, S.; Karhanek, M.; Akhras, M.; Davis, R. W.; Pourmand, N. PathogenMIPer: a tool for the design of molecular inversion probes to detect multiple pathogens. *BMC Bioinformatics* **2006**, *7*, No. 500.

(31) Hu, M.; Yang, B.; Cheng, Y.; Radda, J. S. D.; Chen, Y.; Liu, M.; Wang, S. ProbeDealer is a convenient tool for designing probes for highly multiplexed fluorescence in situ hybridization. *Sci. Rep.* **2020**, *10*, No. 22031.

(32) Khandelwal, G.; Bhyravabhotla, J. A phenomenological model for predicting melting temperatures of DNA sequences. *PLoS One* **2010**, *5*, No. e12433.

(33) Trotta, E. On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One* **2014**, *9*, No. e113380.

(34) Natarajan, A.; Beena, P. M.; Devnikar, A. V.; Mali, S. A systemic review on tuberculosis. *Indian J. Tuberc.* **2020**, *67*, 295−311.

(35) Paton, N. I.; Borand, L.; Benedicto, J.; Kyi, M. M.; Mahmud, A. M.; Norazmi, M. N.; Sharma, N.; Chuchottaworn, C.; Huang, Y.; Kaswandani, N.; Van, H. L.; Lui, G. C. Y.; Mao, T. E. Diagnosis and management of latent tuberculosis infection in Asia: Review of current status and challenges. *Int. J. Infect. Dis.* **2019**, *87*, 21−29.

(36) World Health Organization (WHO), *Global Tuberculosis Report 2021*. World Health Organization: Geneva, 2021.

(37) Reddy, T. B. K.; Riley, R.; Wymore, F.; Montgomery, P.; DeCaprio, D.; Engels, R.; Gellesch, M.; Hubble, J.; Jen, D.; Jin, H.; Koehrsen, M.; Larson, L.; Mao, M.; Nitzberg, M.; Sisk, P.; Stolte, C.; Weiner, B.; White, J.; Zachariah, Z. K.; Sherlock, G.; Galagan, J. E.; Ball, C. A.; Schoolnik, G. K. TB database: an integrated platform for tuberculosis research. *Nucleic Acids Res.* **2009**, *37*, D499−508.

(38) Rosenthal, A.; Gabrielian, A.; Engle, E.; Hurt, D. E.; Alexandru, S.; Crudu, V.; Sergueev, E.; Kirichenko, V.; Lapitskii, V.; Snezhko, E.; Kovalev, V.; Astrovko, A.; Skrahina, A.; Taaffe, J.; Harris, M.; Long, A.; Wollenberg, K.; Akhundova, I.; Ismayilova, S.; Skrahin, A.; Mammadbayov, E.; Gadirova, H.; Abuzarov, R.; Seyfaddinova, M.; Avaliani, Z.; Strambu, I.; Zaharia, D.; Muntean, A.; Ghita, E.; Bogdan, M.; Mindru, R.; Spinu, V.; Sora, A.; Ene, C.; Vashakidze, S.; Shubladze, N.; Nanava, U.; Tuzikov, A.; Tartakovsky, M. The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. *J. Clin. Microbiol.* **2017**, *55*, 3267−3282.

(39) Bhasuran, B. Combining Literature Mining and Machine Learning for Predicting Biomedical Discoveries. *Methods Mol. Biol.* **2022**, *2496*, 123−140.

(40) Joffroy, B.; Uca, Y. O.; Prešern, D.; Doye, J. P. K.; Schmidt, T. L. Rolling circle amplification shows a sinusoidal template length-dependent amplification bias. *Nucleic Acids Res.* **2018**, *46*, 538−545.

(41) Pearson, W. R. Finding Protein and Nucleotide Similarities with FASTA. *Curr. Protoc. Bioinf.* **2016**, *53*, 3.9.1−3.9.25.

(42) Maglott, D.; Ostell, J.; Pruitt, K. D.; Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **2011**, *39*, D52−D57.

(43) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.

(44) Sievers, F.; Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **2014**, *1079*, 105−116.

(45) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276−277.

(46) Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022−3027.

(47) Tamura, K.; Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **1993**, *10*, 512−526.

(48) Nie, L.; Cai, S.; Shao, J.; Chen, J. Toll-Like Receptors, Associated Biological Roles, and Signaling Networks in Non-Mammals. *Front. Immunol.* **2018**, *9*, No. 1523.

(49) Di Franco, A.; Poujol, R.; Baurain, D.; Philippe, H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* **2019**, *19*, No. 21.

(50) Gil, N.; Fiser, A. The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis. *Bioinformatics* **2019**, *35*, 12−19.