LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-JRNL-491362

Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments

D. Hysom, P. Naraghi-Arani, P. Williams, S. Gardner

July 26, 2011

Nucleic Acids Research

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments

David A. Hysom[1], Pejman Naraghi-Arani[2], Peter L. Williams[3], Shea N. Gardner[1,2,*]

[1]Computations, Lawrence Livermore National Laboratory, Livermore CA 94550

[2]Global Security, Lawrence Livermore National Laboratory, Livermore CA 94550

[3]Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598

*Author for correspondence, gardner26@LLNL.gov

Phone: 925-422-4317

Fax: 925-422-6736

Author emails: hysom@llnl.gov, pnaraghi@llnl.gov, plwilliams@lbl.gov

Keywords: primer design, degenerate primers, multiplex primers, TaqMan signature, Rift Valley fever virus, software, Primux

4796 words + 3 figures + 7 tables
Estimated length=8.7 printed pages

Abstract

Primux is a new software package for selecting multiplex compatible, degenerate primers and probes to detect diverse targets such as viruses. It requires no multiple sequence alignment, hence it scales well for large target sets and saves user effort from curating sequences into alignable groups. It improves on the previous MPP software predicting non-degenerate primer sets using k-mer algorithms [1] by adding the capability to predict degenerate primers as well as probes suitable for TaqMan or other primer/probe triplet assay formats, or simply probes for microarray or other single-oligo assay formats. Primux employs suffix array methods for memory efficient calculations on oligos 10-~100 nt in length. Primers and probes for each segment of Rift Valley fever virus were designed using Primux, and lab testing showed excellent sensitivity and species specificity.

Introduction

Viral species detection can be difficult when high levels of strain variation have evolved. Substantial investment is required to obtain adequate sequence data to represent the known diversity and to design conserved, species specific signatures [2, 3]. Many published PCR-based signatures are not robust, and in a computational analysis of dozens of published signatures, over 60% of the viral signatures analyzed failed to detect all desired targets based on available sequences [4]. While methods like sequencing and microarrays can overcome this problem and detect a wide range of viruses of diverse strains and unanticipated species [5], PCR-based assays aimed at detecting one or several species are faster and cheaper if screening is limited to a handful of candidates. As the amount of sequence data skyrockets with advances in sequencing technology, signature design software must scale up to keep pace.

Designing primers for sets of diverse target sequences typically starts with Multiple Sequence Alignment (MSA) to identify the most conserved regions for primer design. Obtaining a good MSA can be a bottleneck in terms of compute time and memory, and alignments of many divergent sequences can be of questionable quality. Sequence rearrangements also thwart some alignment tools. In previous work [1], many degenerate primer design software packages, all requiring an MSA as input, were tested. Most of these failed to complete for the viruses tested. Only Greene SCPrimer [6] and HYDEN [7] would complete for the two smallest target sets, Norwalk virus and foot and mouth disease virus (FMDV). The novel k-mer approach to non-degenerate primer prediction used by the Multiplex Primer Prediction software (MPP) presented in [1] completed for both the smallest target sets and other much larger and more diverse target sets. These larger sets included primer set prediction for all available genomes for each of 79 viral families. The authors of the PAMPs software [8] claim that it is faster than PT_MIPS [9] and predicts smaller primer sets to detect all targets, but the size of the total input sequence file is limited to 2000 nucleotides, severely limiting the scalability and application of the of the software and precluding analysis of large target sets of multiple genes or genomes.

MPP computes primers and/or probes by operating on sets of k-mers (oligos of length k) shared by multiple sequences, and hence bypasses the MSA phase. However, MPP requires that the oligos be exact matches, i.e, do not contain degenerate (mismatching) bases. The new Primux software improves on MPP in that it permits a small number of mismatches in the primers and probes. Without degenerate bases, the use of shorter (less than 18 base) primers is sometimes necessary in order to reduce the final primer set to a manageable size, that is, a set small enough to meet wet-lab multiplex assay constraints. The advantage of allowing degenerate bases is that primers of at least 18 bases can often still be found that are conserved across multiple targets without needing to reduce primer length. Maintaining primer length of 18 bases or longer, with higher Tm, results in better primer binding to target for improved sensitivity and specificity compared to that of shorter primers.

Primux uses the following approach. After initial k-mer enumeration from all target sequences using storage and memory-efficient suffix array methods (McIlroy, T.M. and McIlrow, M.D., Sarray, a collection of Suffix-array functions. http://www.cs.dartmouth.edu/~doug/sarray/ Copyright (C) Lucent Technologies) -- k-mers are ranked by conservation among the targets. The most conserved k-mers are filtered for optimal primer or probe characteristics such as Tm, GC%, avoidance of homopolymer runs, homodimers, hairpin folds, and repetitive sequence. Conservation is then re-calculated allowing up to a user-specified number of mismatches.

Several alternative greedy algorithms are then employed to select alternative sets of k-mer forward/reverse primer pairs. A “min” algorithm picks the most conserved primer pairs that will detect the *outlying, least conserved targets* first. A “max” algorithm picks the most conserved primer pairs that will detect the *most conserved targets first*. A “combo” algorithm that does a combination of “min” and “max”. The algorithm that yields the fewest primer pairs that detect all targets differs for various target sets, but “max” usually performs best in the tests we have performed. In addition, to achieve multiplex compatible primer sets, if nearest neighbor hybridization free energy calculations predict that a primer might dimerize with other primers already selected, it is dropped and the next most conserved primer is considered. Because there may be alternative, equivalent primers that can be selected at each step, different primer sets can be designed, and the user can tell the software how many sets are desired. This is advantageous since, if one set fails in lab testing, there are back-ups (In practice, we typically compute three alternative sets for each of the three algorithms.) The alternative sets may differ in size since the algorithms follow a greedy heuristic and do not compute a true optimum, as the minimal set degenerate primer selection problem has been proven to be NP complete [7]. Many of the key features of Primux that distinguish it from previous primer prediction softwares are listed in Figure 1.

From the amplicons generated by a primer set, Primux can design conserved probes to detect all targets with the fewest probes, or genotyping probes to maximize the target discrimination. The resulting primer and probe triplets are suitable for TaqMan, Luminex/BioPlex, or other similar assays. The probe selection process starts by computing all the k-mers on the amplicons and ranking by conservation, filtering by Tm, GC%, etc. as for the primers, and recalculating conservation allowing mismatches. For the fewest probes to detect all targets, the most conserved probes are selected for each target. For genotyping probes, the least conserved probes are selected for each target.

Primer/probe triplets with both conserved and genotyping probes were predicted for each (L, M, S) segment of Rift Valley fever virus. 54, 60, and 84 sequences were available for the L, M, and S segments, respectively. These were tested in the lab against target, background, and near neighbor viruses.

Methods

Figure 2 charts the approach employed by Primux. The Primux pipeline is implemented as a series of modules (i.e., scripts and executables) whose inputs and outputs are tied together via the file system. Compute intensive modules are coded in C ++. Other modules (primarily those involved in parsing and invoking third-party executable) are coded in python or perl. The Primux pipeline for computing forward/reverse primer pairs is run by invoking a python script, **findPrimers.py,** which takes a single command line parameter, the name of an options file.

The options file specifies all settings relevant to an experiment. These include the name of the input fasta-formatted file; settings to be passed to UnaFold [10]; the number of permitted degenerate base pairs in the primers; etc. Options files follow a simple text-based format, e.g;

```
#this is a comment
# directory in which results are stored
-dir=results

#min, max and increment for computing kmers
-min_kmer_len=18
-max_kmer_len=22
-kmer_len_inc=1
```

Third party codes

In addition to in-house code, Primux uses the third party codes: ssarray.tar (http://www.cs.dartmouth.edu/~doug/sarray/), UNAFold [10], Vmatch (Stephan Kurtz: The Vmatch large scale sequence analysis software, http://www.vmatch.de).

Terminology

A **k-mer** is a relatively short sub-sequence of a genome or DNA fragment. More pedantically, a k-mer is a sequence of characters that represents the sequence of nucleotides (nt) in an oligo. Given a set of DNA sequences, **TopN** refers to the set of the *N* most frequently occurring k-mers. Although a k-mer may occur multiple times in a single DNA sequence, for the purposes of this work we tabulate TopN on the basis: a k-mer appears at least once in a DNA sequence (i.e., multiple occurrences are not significant). More expansively, a TopN set is computed as follows. Given a set of genomes and a value *N*, we compute the most frequently occurring kmers across all genomes.

BottomN is analogous to TopN, but refers to the set of least frequently occurring k-mers. TopN sets are used when computing a minimal set of PCR signature and probes, while BottomN sets are used when computing genotyping probes.

A **SuperSet** is a data structure that contains a k-mer, along with a listing of where the k-mer occurs in a given set of DNA sequences. A SuperSet also lists the locations of the degeneracies (if any) for each occurrence. Additionally, a SuperSet contains a field indicating if its k-mer is to be used in the forward, reverse complement, or either direction (f/r/e). To make this more concrete, an example SuperSet contains the following information:

#the kmer, with no degeneracies
ACTGTCTAGCA
#the kmer occurs in sequence 3, with the degenerate in position one filled by 'C'
CCTGTCTAGCA :: 3 :: 1, C;
#the kmer occurs in sequence 5, with degenerate in position 1 filled by 'T' and degenerate position 6 filled by 'A':
TCTGT**A**TAGCA :: 5 :: 1, T ; 6, A
#the kmer occurs in sequence 6 with no degeneracies
ACTGTCTAGCA :: 6

A **PrimerPair** is a data structure that contains two SuperSets such that the k-mers in the SuperSets represent a PCR signature. This means that the distance between the two oligos is between the minimum and maximum amplicon length; and that both oligos have passed all filters (dimerization, melting temperature, etc; see below).

Module and Pipeline Descriptions

Figure 3 contains a flow diagram for Primux's primer finding pipeline. The probe finding pipelines use many of the same modules.

The **SuffixArray** module computes the TopN and/or BottomN k-mers for a set of genomes and a given *k* (or range of *k*). This module uses the C-based Suffix-array code sarray as its compute-intensive kernel. The output set of kmers, which are written to a fasta-formatted file, are non-degenerate. We call the output the *k-mer query file*.

The **Vmatch** module uses S. Kurtz's executables as its computational kernel. It takes as input the k-mer query file and an integer that specifies the maximum permitted number of degenerate base pairs, and outputs a file formatted as described at **http://www.vmatch.de**.

The **pvmo** module parses the output from Vmatch and constructs and writes to file a set of SuperSets. We refer to this as the *k-mer SuperSet file*.

The **findPotentialPartners** module takes as input the SuperSet file from the previous module. For each position of every k-mer in every SuperSet file it computes a set of ***potential partner*** k-mers. A potential partner is a k-mer that is within the required distance (between the desired minimum and maximum amplicon lengths, as specified in the options file) such that the k-mer and its potential partner form a pair of primers. The potential partners are generated using the **SuffixArray** module, and query files are written as described above. We search in both before and after directions to find potential

partners. Hence we end up with two query files, the *before query file* and *after query file*. As above, the before and after query files are processed using **Vmatch**, and parsed to form sets of *before* and *after SuperSets*.

The **computePrimerPairCandidates** module takes as input the *k-mer SuperSet file* and the *before* and *after SuperSet files* and outputs a set of PrimerPairs. Instead of outputting all possible PrimerPairs, we down-select to ensure that we output a relatively small number (perhaps thousands), that the selected PrimerPairs cover the largest possible number of genomes, but that every genome is covered by a minimal number of PrimerPairs.

The **findPrimerSet** modules take as input a set of PrimerPairs and output one or more (ideally minimal) set(s) of primers suitable for PCR-based amplification of the target genomes. Each PrimerPair in the input set has already been determined to be a valid PCR signature. The goal of these modules, then, is to select a set of PrimerPairs such that (i) every sequence in the target set can be amplified by at least one forward/reverse complement pair of primers; (ii) the resulting set of primers is minimal.

As noted above, we have devised several heuristic algorithms that accomplish this aim. Our ***min*** approach picks a set of k-mer forward/reverse pairs based on the criterion: select primers that will detect the *least* conserved targets first. Our ***max*** algorithm picks a set of k-mer forward/reverse pairs based on the criterion: select primers that will detect the *most* conserved targets first. Our ***combo*** algorithm utilizes both min and max strategies. First, one primer pair is selected that covers the most conserved targets; this is the max phase. Then, additional pairs are added using the min strategy, until all targets are covered by at least one primer pair.

Each algorithm also takes as input a number of requested iterations. For each iteration, the algorithms work as described above, but exclude from consideration any PrimerPairs that were previously selected.

After a set of primers has been found to detect all targets, it is fixed to remove any extra primers that result in redundant coverage of any targets, which can occur due to the greedy algorithm used. The final sets of primers are reported in files beginning in the word "fixed". Redundant coverage can occur, for example, if the 1st selected primer-pair covers sequences 1-4; then 2nd covers 5, but also covers 1 and 2; the 3rd covers 6, but also covers 3 and 4. So now the 1st primer is not needed.

Probes

After computing sets of primer pairs per above, we extract all amplicons (minus the forward and reverse primers) from the target set into a fasta-formatted file. This file is input to the **findConservedProbes.py** script, which is identical to the findPrimers.py pipeline, through computation of the k-mer SuperSet file. It then applies a max algorithm that computes one or more minimal sets of k-mers (the probe set), such that each sequence in the input file contains a probe from the set.

The **findDiscrininatingProbes.py** script attempts to find sets of strain discriminating probes. That is, given a set of amplicons from a given primer set, it attempts to generate a set of probes to discriminate the targets to the maximum level allowed by the sequence diversity of those amplicons. This pipeline is similar to findConservedProbes.py, however, instead of working with the TopN k-mers it

works with the BottomN sets.

Filters

At various stages in the pipeline filters are applied to eliminate k-mers that violate constraints specified in the options file. Filters include:

- **UNAFold**: does a k-mer adhere to the specified minimum and maximum melting temperatures, hairpin, and homdimerization constraints?

- **GC percent**: does a k-mer's GC percent fall within the specified minimum and maximum constraint?

- **entropy**: is a k-mer's entropy above the specified minimum? (Entropy is computed as $E = \sum_{t:f_t \neq 0} - f_t \log_2 f_t$ where $f_t = f_{AAA}, \ldots, f_{TTT}$ are the frequencies of each of the 64 possible trimers in the k-mer sequence, calculated as the number of occurrences of each trimer divided by the total number of trimers in the sequence. The sum is over the trimers t with $f_t>0$. The entropy filter eliminates repetitive, low complexity sequence that tends to function poorly as primers.

- **min_dist_mm_to_3prime_end**: a k-mer position in a SuperSet is deleted if it contains a degenerate base pair that is within min_dist_to_3prime-end base pairs of the 3prime end.

- **maxPolyX**: the maximum permitted number of identical sequential base pairs in a k-mer. By example, if maxPolyX=3, the kmer ACTTT would be rejected due to the sub sequence 'TTT.'

- **dimerization:** when selecting the primer sets (per findPrimerSet above), does each potential primer fall within the (UNAFold calculated) dimerization constraints for all previously selected primers?

Computational Examples

Signatures for the L, M, and S segments of RVF virus were designed from all available complete segment sequences. The sequence identifiers and parameter options used as input to Primux are provided as supplemental information. We required primers to be 18-22 nt long, Tm=60-65°C, up to 3 degenerate bases no closer than 3 nt from the 3' end, no more than 3 bases in a homopolymer repeat, GC% of 20-80%, and produce amplicons 80-250 bp. We used Primux to design sets of conserved, multiplex compatible degenerate primers and conserved probes for each segment to detect all sequences of that segment. Table 1 summarizes the numbers of sequences, primers, and probes.

Rift Valley Fever virus signature testing

Virus Culture

All culture was conducted at the Biosafety Level 4 (BSL-4) facilities at the University of Texas Medical Branch and titered viral RNA kindly provided by Dr. Alexander Freiberg's group. The RVFV signatures designed by Primux as well as those designed using more traditional bioinformatics tools (see below) were tested on total nucleic acid extracts generated from RVFV grown on Vero C1008 [Vero 76, clone E6; Vero E6] cells (ATCC # CRL-1586) using standard viral culture methods (RVFV Reference 1) and virus generation determined by cytopathic effect (CPE). RVFV titers were determined by standard plaque assays on Vero E6 or Vero CCL-81 (ATCC # CCL-81) cells [11,12,13].

Briefly, cells were cultured in T25 flasks under 10% media at 37C with 5% CO2 until 90% confluent. At total of 1ml of viral inoculum (1:200 dilution of virus stock in 2% growth media) or negative control (1 ml of 2% growth media with no virus) was used as inoculums and incubated for 1 hour at 37C, 5%CO2. Volume of media was brought to 15ml total with 2% growth media and incubated for 2-5 days until 70% CPE was observed by inverted microscope examination. Viral titers were determined by plaque assay using standard procedures [12]. Cultured viral supernatant was clarified by low speed centrifugation and diluted 1:3 with Trizol LS reagent (Life Technologies, Carlsbad, CA). Total RNA was extracted from viral samples and tested by attempted culture for 14 days to verify the absence of viable virus prior to viral samples being removed from BSL-4 and shipped to LLNL under Trizol-LS.

Bioinformatics design of standard (non-degenerate) RVFV Assay Oligonucleotide signatures

All candidate (primer–probe triplets designed to target specific regions of a genome) signatures were designed using a whole-genome approach [14]. Briefly, all available complete genomes of different strains of RVFV were compared using multiple genome alignment programs. All available complete or partial RVFV genomes were computationally examined to identify sequence regions that were conserved among all sequenced isolates of RVFV but also able to uniquely identify RVFV when compared against all available genomic sequence data. A final *in-silico* screening was performed to verify that signatures were predicted to detect the RVFV strains and not to detect any non-pathogen targets. The conserved/unique sequence information was used to develop candidate signatures that met Real-Time RT-PCR assay chemistry requirements.

PCR Primers and Probes

Oligonucleotide primers and probes were purchased as lyophilized pellets from Biosearch Technologies, Inc. (Novato, CA). Purification of each lot was assessed by mass spectrometry. TaqMan probes were designed with a 5' Fam and a 3' Black Hole Quencher molecule. Upon receipt, oligos were reconstituted in sterile 1X Tris-EDTA (TE) Buffer (10mM Tris-Cl, 1mM EDTA, pH 8.0, Teknova, Hollister, CA) to a concentration of 100 mM. Working stocks were made by diluting primers and probes to a concentration of 10 mM with TE Buffer. The primer and probe working stocks were stored at 4°C. Unused 100mM primers and probes were stored at -80°C.

RT-PCR

All assay reactions were carried out on 96 well FAST PCR plates (Applied Biosystems, Foster City, CA) in a total volume of 25 µl (20 µl master mix plus 5 µl sample) optimized for Real-Time RT-PCR. A volume of 20 µl Real-Time RT-PCR master mix (AgPath-ID™ One-Step RT-PCR Kit, Cat#4387391, Applied Biosystems, Foster City, CA [Life Technologies, Carlsbad, CA]) contains: 12.5µl 2X RT-PCR Buffer Mix, 1 µl

25X RT-PCR Enzyme Mix, 1.0 µl primer mix (0.4 µM each forward and reverse primer final concentration), 0.5 µl TaqMan probe (0.2 µM final concentration), 5.0 µl PCR water (Teknova Inc), and varying amounts of template RNA. Each plate contained 3 negative (no template controls) and 3 positive controls containing 1000 copies of Alien-armored RNA for each signature. Alien armored RNA is heat lysed prior to addition to the mastermix by diluting the Alien-armored RNA to 200 copies/µl in water and placing in a heat block for 3 minutes at 75ºC. The Alien-armored RNA (XenoRNA-01, Ambion, Austin, TX) is a proprietary 1070 nucleotide RNA transcript consisting of unique nucleotide sequences that possess no significant homology to the current annotated sequences in commonly used sequence databases.

Reactions were carried out on ABI 7500 thermal cyclers (Applied Biosystems, Foster City, CA) under the following quantitative Reverse Transcriptase Fast thermal cycling conditions: 45ºC for 10 minutes for cDNA synthesis, followed by 95ºC for 10 minutes for inactivation of the reverse transcriptase, activation of 25X RT-PCR Enzyme Mix, and denaturation of the RNA/cDNA hybrid; followed by amplification at 40 cycles of 97ºC for 2 seconds and 60ºC for 30 seconds. Data acquisition was performed at the annealing step.

Primer/Probe Sequences: Oligonucleotide sequences for the assays used in this study are listed in Tables 2 and 3 below. Table 2 lists oligonucleotides with incorporated degeneracies designed using the Primux system while Table 3 lists oligonucleotides with no degeneracies designed using our traditional signature prediction software. All signatures were designed to have optimal annealing temperatures of 60°C to 65°C enabling testing of all signatures using a single thermal cycling protocol as listed above.

Nucleic Acids Extraction

All template RNAs were extracted with three times the volume of Trizol (Invitrogen, Carlsbad, CA), mixed, and incubated 15 min at 25°C. One-fifth of the total volume of chloroform was added, mixed, incubated 15 min at 25°C and centrifuged at 3000 x g 15 min at 4°C. To the aqueous layer, 70% volume of 100% isopropyl alcohol was added. The sample was mixed, incubated 10 min at 25° C, and then centrifuged at max g for 10 min at 4°C. The pellet was washed with 70% ethanol and then centrifuged at max g for 10 min at 4°C, air dried briefly at room temperature, dissolved in RNase-free, DEPC treated water (Ambion, Austin, TX) and stored at -80 ° C until needed.

The concentration of extracted RNA was determined for the samples using a ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE). Final titer amounts were determined using the titer information provided by UTMB for the total nucleic acid extracts from virus culture stocks provided.

Titrations to Determine Limits of Detection

All signatures that were down-selected were tested using eight-fold serial dilutions of extract RNA from six titered and one untitered Rift Valley Fever strains in the previously described Real-Time rt- PCR format. The total amount of titered and untitered template added to each reaction ranged from 1000-0.0001pfu and 100-0.00001 pg total RNA. The reactions were performed in triplicate. The limit of detection (LOD) for each signature was defined as the minimum amount of target template required to generate a Ct value equal to or smaller than 35.

Results

Two sets of qRT-PCR signatures utilizing the "TaqMan" chemistry were designed using either our traditional approach for such signature design or the Primux system. Both sets of signatures were tested on the same set of RVFV target and near-neighbor nucleic acid extracts. The degenerate primer/probe sets were attempted in TaqMan format with multiple combinations of each primer set with various probes. The results shown reflect those primer/probe set combinations that generated no cross-reactivity with background samples or near-neighbor samples and which demonstrated the ability to detect all target strains.

Specificity Testing on Target, Near Neighbor and Background Nucleic Acids

A collection of purified RNA samples was used for signature testing and down-selection. Part of this collection was generated and generously provided by the Freiburg laboratory at the Center for Biodefense and Emerging Infectious Diseases at the University of Texas Medical Branch (UTMB, Galveston, TX) and included RNA extracts from a total of six titered targets, one un-titered target (Table 4), and 16 genetic near neighbors of Rift Valley Fever (RVF) virus strains (Table 5). Near-Neighbor strains are defined as genetically closely-related organism that are phylogenetically similar (related) to and yet distinct from our target organism. Each signature was screened in triplicate against the target and near neighbor RNA extracts. All initial target reactions were performed using an amount of total RNA extract that was equivalent to 1000 pfu total RNA from titered viral samples. For the untitered MP-12 (vaccine) strain, 100 pg of total RNA extract was used. All near neighbor reactions were performed using 1ng total RNA extracts from viral culture samples generated as RVFV. Each signature was additionally screened against a collection of "background" nucleic acids. The background panel comprises nucleic acid extracts from selected variety of mammalian, avian, and arthropod cell lines (Table 6). This panel is designed to challenge the specificity of each signature, ensuring that the signatures react only with target nucleic acids and not the nucleic acids from various potential viral hosts. RNA extracts were quantitated on an ND-1000 Spectrophotometer (Nanodrop, Wilmington, VT). Each signature was screened in triplicate against background cell lines. Background reactions were performed using 200 pg total RNA.

Table 7 lists the primer "S, M, or L" and the probe "C" combination from each set of signatures listed in Tables 2 and 3 that resulted in no cross-reactivity with any non-target template and were able to detect all target strains. Results are reported for Limit of Detection studies conducted in triplicate reactions for each signature/target combination. Cycle threshold cutoffs were 35 (ie: any Ct greater than 35 was defined as a detection failure. Thus, for each RVFV strain tested, the LOD reported in Table OOO is the template concentration at which all three replicates of the RT-PCR reaction generated a Ct value equal to or below 35.

The results in Table 7 demonstrate that all degenerate signatures performed at least as well as the non-degenerate signatures for detection of SA-75, KEN58, ZH501, and MP-12 strains of RVFV. The non-degenerate signatures were able to detect the ZH548 strain of RVFV better than two of the degenerate signatures (L1C2 and L1C3), but the performance of each of the other 4 degenerate signatures was essentially equivalent to that of the non-degenerate signatures. Finally, the degenerate signatures tested performed better than the non-degenerate signatures for detection of Mauritania OS-1, and SA51 strains of RVFV by as much as 6 logs (comparing any non-degenerate signature with L1C1 or L1C2).

Discussion

We describe Primux, a k-mer based approach to designing primers and probes for detecting diverse sets of target sequences. No multiple sequence alignment is required. Instead, a greedy algorithm based on k-mer analysis with suffix arrays identifies conserved, degenerate k-mers that meet primer specifications (Tm, etc.) and which can be combined in multiplex to amplify at least one fragment from each of the target sequences. Additional k-mer based codes are described to predict conserved or discriminating sets of probes for a set of input sequences. These input sequences can be the amplicons produced by Primux-designed primers, resulting in primer and probe triplets for use in assay formats such as TaqMan or Luminex.

Labaratory tests comparing the performance of degenerate primers and probes designed using the Primux software to signatures designed utilizing non-degenerate bases on multiple background and near-neighbor templates demonstrated that the degenerate signatures designed by Primux performed as well as non-degenerate signatures in specificity tests (ie: no cross-reactivity was observed). Furthermore, multiple degenerate signatures performed as well as or better than non-degenerate signatures (by up to 6 logs) designed by more traditional software approaches in sensitivity assessments using multiple RVFV strains. The Primux software is available as open source from http://sourceforge.net/projects/primux.

Author contributions

DH and PW built the software, PN tested the RVF signatures in the lab, and SG lead the work, tested the software, and designed the RVF signatures. DH, PN, and SG wrote the paper.

Funding

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was supported by Laboratory Directed Research and Development grant number 08-SI-002 from Lawrence Livermore National Laboratory (https://www.llnl.gov/).

Acknowledgments

We would like to thank Drs. Alexander Freiberg & Robert Tesh and Ms. Colette Pitze for generation of RVFV target and genetic near-neighbor templates used for this study.

References

1. Gardner, S.N., Hiddessen, A.L., Williams, P.L., Hara, C., Wagner, M.C., and Colston, B.W., Jr. (2009) Multiplex primer prediction software for divergent targets. *Nucl Acids Res*, **37**(19):6291-6304.
2. Gardner, S.N., Kuczmarski, T.A., Vitalis, E.A., and Slezak, T.R. (2003) Limitations of TaqMan PCR for Detecting Divergent Viral Pathogens Illustrated by Hepatitis A, B, C, and E Viruses and Human Immunodeficiency Virus. *J Clin Microbiol* , **41**(6):2417-2427.
3. Gardner, S.N., Lam, M.W., Mulakken, N.J., Torres, C.L., Smith, J.R., and Slezak, T.R. (2004) Sequencing Needs for Viral Diagnostics. *J Clin Microbiol*, **42**(12):5472-5476.
4. Lemmon, G.H. and Gardner, S.N. (2008) Predicting the sensitivity and specificity of published real-time PCR assays. *Ann Clin Microbiol Antimicrob*, **7**:18. doi: 10.1186/1476-0711-1187-1118.
5. Victoria, J.G., Wang, C., Jones, M.S., Jaing, C., McLoughlin, K., Gardner, S., and Delwart, E.L.(2010) Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. *J Virol*, JVI.02690-02609.
6. Jabado, O.J., Palacios, G., Kapoor, V., Hui, J., Renwick, N., Zha,i J., Briese, T., Lipkin, W.I.(2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucl Acids Res,* **34**(22):6605-6611.
7. Linhart, C. and Shamir, R. (2002)The degenerate primer design problem. *Bioinformatics*, **18**(suppl_1):S172-181.
8. Najafabadi, H., Torabi, N. and Chamankhah, M. (2008)Designing multiple degenerate primers via consecutive pairwise alignments. *BMC Bioinformatics*, **9**(1):55.
9. Souvenir, R., Buhler, J., Stormo, G. and Zhang, W. (2003) Selecting degenerate multiplex PCR primers. *Proceedings of the 3rd Workshop on Algorithms in Bioinformatics (WABI 2003)*, 512-526.
10. Markham, N.R. and Zuker, M.(2008) UNAFold: software for nucleic acid folding and hybridization. In Keith, J.M. (ed.), *Bioinformatics, Volume II Structure, Functions and Applications, number 453 in Methods in Molecular Biology.* Humana Press, Totowa, NJ, Vol II, pp. 3-31.
11. Wagner, E.K. and Martinez, J.H. (2004) *Basic Virology* 2nd Edition, Blackwell Publishing, pp136-138.
12. Mo, J.B., Lambert, R.D. and Lupton, H.W. (1981) Plaque Assay for Ebola virus. J.*Clin. Microbiol*., **13**(4):791-793.
13. O'Mahony, J., O'Donoghue, M., Morgan, J.G. and Hill, C. (2000). Rotavirus survival and stability in foods as determined by an optimized plaque assay procedure. *Int. J. Food Microbiol.*, **61**(2-3):177-85.
14. Slezak, T., Kuczmarski, T., Ott, L., Torres, C., Medeiros, D., Smith, J., Truitt, B., Mulakken, N., Lam, M., Vitalis, E., Zemla, A., Zhou, C. E., and Gardner, S. N. (2003) Comparative genomics tools applied to bioterrorism defense. *Briefings in Bioinformatics*, **4**: 133-149.

Figure Legends

Figure 1: Text box listing important aspects of Primux

Figure 2: Diagram of the Primux approach

Figure 3: Diagram of the Primux software workflow.

Primux's Novel Approach

- No multiple sequence alignment of target sequences required
  - ideal for RNA viruses that align poorly
  - Ideal for extremely large viral target sets with hundreds of sequences
- Employs recent advances in string processing algorithms and data structures (computer science research) to efficiently find conserved primer candidates and degenerate primers
- Applies best available thermodynamic predictions to computationally avoid likely causes of multiplex failure like primer dimers, hairpins, and incompatible Tm's
- Iterative steps to build a multiplex primer set ensure that ALL targets are detected by final set
- User-specified primer characteristics can be tailored for any desired reaction protocol
- User option to find multiple, alternative, non-overlapping primer set solutions

Figure 1

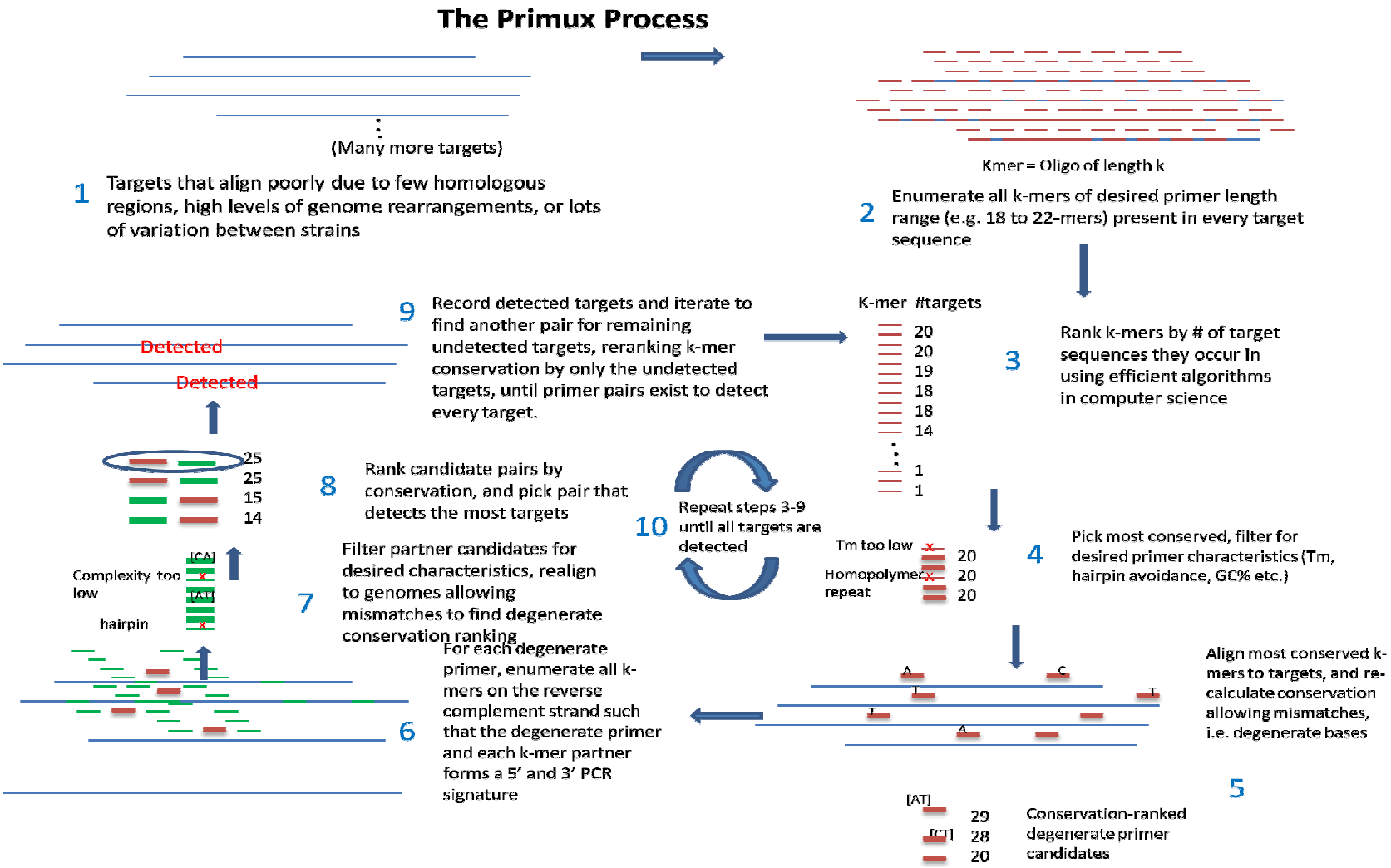
The Primux Process
(Many more targets)
1 Targets that align poorly due to few homologous regions, high levels of genome rearrangements, or lots of variation between strains
Kmer = Oligo of length k
2 Enumerate all k-mers of desired primer length range (e.g. 18 to 22-mers) present in every target sequence
K-mer #targets
20
20
19
18
18
14
1
1
3 Rank k-mers by # of target sequences they occur In using efficient algorithms in computer science
Tm too low
Homopolymer repeat
20
20
20
4 Pick most conserved, filter for desired primer characteristics (Tm, hairpin avoidance, GC% etc.)
A
C
T
T
A
5 Align most conserved k-mers to targets, and re-calculate conservation allowing mismatches, i.e. degenerate bases
[AT]
[CT]
29
28
20
Conservation-ranked degenerate primer candidates
6 For each degenerate primer, enumerate all k-mers on the reverse complement strand such that the degenerate primer and each k-mer partner forms a 5' and 3' PCR signature
7 Filter partner candidates for desired characteristics, realign to genomes allowing mismatches to find degenerate conservation ranking
Complexity too low
hairpin
[CA]
[AT]
8 Rank candidate pairs by conservation, and pick pair that detects the most targets
25
25
15
14
9 Record detected targets and iterate to find another pair for remaining undetected targets, reranking k-mer conservation by only the undetected targets, until primer pairs exist to detect every target.
Detected
Detected
10 Repeat steps 3-9 until all targets are detected

Figure 2

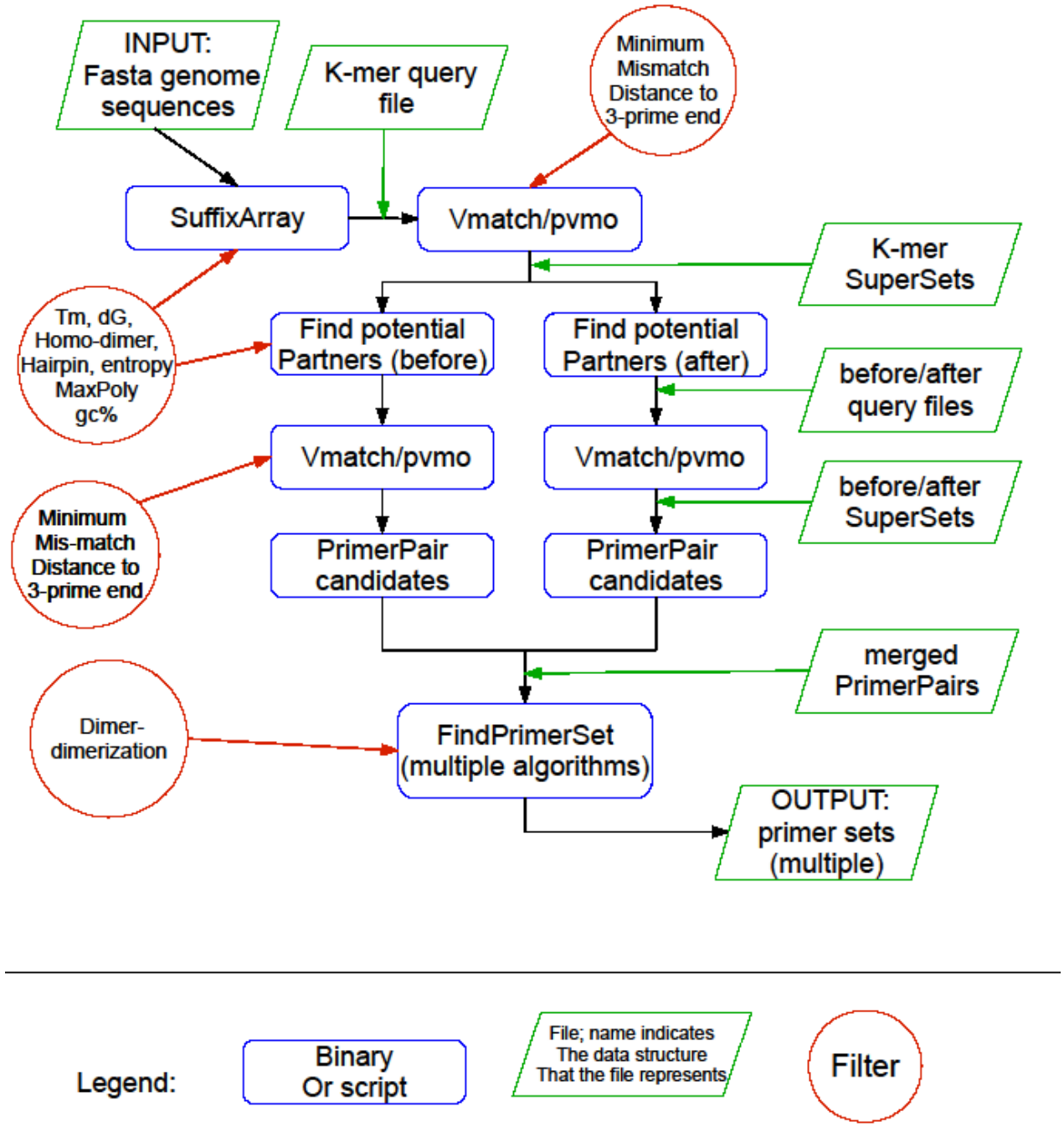
INPUT:
Fasta genome
sequences
K-mer query
file
Minimum
Mismatch
Distance to
3-prime end
SuffixArray
Vmatch/pvmo
K-mer
SuperSets
Tm, dG,
Homo-dimer,
Hairpin, entropy
MaxPoly
gc%
Find potential
Partners (before)
Find potential
Partners (after)
before/after
query files
Vmatch/pvmo
Vmatch/pvmo
before/after
SuperSets
Minimum
Mis-match
Distance to
3-prime end
PrimerPair
candidates
PrimerPair
candidates
merged
PrimerPairs
Dimer-
dimerization
FindPrimerSet
(multiple algorithms)
OUTPUT:
primer sets
(multiple)
Legend:
Binary
Or script
File; name indicates
The data structure
That the file represents
Filter

Figure 3

Table 1: Summary of RVF target sequence, primer and probe counts.

Segment	Number of target sequences	Number of primers	Number of conserved probes
S	84	2	2
M	60	4	2
L	54	4	3

Table 2: Rift Valley Fever (RVF) Degenerate Assay Signatures

Signature Name	Forward Primer Sequence	Reverse Primer Sequence	Probe Name	Conserved Probe Sequence
RVF_S	CTTGGCATCCTTCTCCCAG	CAAGCAGTGGACCGCAAT	RVF_S_C1	TTGARCAGTGGGTCCGAGAGTTTG
			RVF_S_C2	TGCTTATCARGGRTTTGATGCCCGTAG
RVF_M1	AGCCATCATTGCTGCYGATG	AGGCTGGAAGGACTGTCA	RVF_M_C1	YCTTACCATRGCAGGTGATGTTGTTCAAGC
			RVF_M_C2	ATTTGAGCCTGARATGCCCTCTGC
RVF_M2	GGAGCATCRTCTAGCCGTTTC	GCATACCCTTTGCCTGGG	RVF_M_C1	YCTTACCATRGCAGGTGATGTTGTTCAAGC
			RVF_M_C2	ATTTGAGCCTGARATGCCCTCTGC
RVF_L1	TGCGTGAGTTTCCCATGA	CTGCCCTGAGATCTGTTCTCAC	RVF_L_C1	CCCCTAAARGTGGTGAACTCAACGATGTT
			RVF_L_C2	ATGCACCYCTTTCATCTCCCCTA
			RVF_L_C3	TCYCCTCTCACATCTARTCCCTGAAGA
RVF_L2	CTRCCTCCCTGGYTGTCC	GCATCATCGTGCATCCTCTCAA	RVF_L_C1	CCCCTAAARGTGGTGAACTCAACGATGTT
			RVF_L_C2	ATGCACCYCTTTCATCTCCCCTA
			RVF_L_C3	TCYCCTCTCACATCTARTCCCTGAAGA

Table 4: Targets used in RVF signature screening.

Targets	
Genus, species	Strain
Phlebovirus, Rift Valley Fever	SA75
Phlebovirus, Rift Valley Fever	Ken 58 (B691)
Phlebovirus, Rift Valley Fever	Mauritania OS-1
Phlebovirus, Rift Valley Fever	ZH548
Phlebovirus, Rift Valley Fever	ZH501
Phlebovirus, Rift Valley Fever	SA51
Phlebovirus, Rift Valley Fever	MP12

Table 5: Near-Neighbors used in RVF signature screening.

Near-Neighbors	
Genus, species	Strain
Nairovirus, Bandia	N/A
Nairovirus, Crimeon Congo Hemorrhagic Fever	Dubai TP9
Nairovirus, Crimeon Congo Hemorrhagic Fever	Matin
Nairovirus, Crimeon Congo Hemorrhagic Fever	Oman 199723179
Nairovirus, Crimeon Congo Hemorrhagic Fever	SPU 97185
Nairovirus, Dugbe	YARU collection 5/27/1971
Nairovirus, Hazara	N/A
Nairovirus, Hughes	DT
Nairovirus, Khashan	Leiv 776P
Nairovirus, Soldado	N/A
Phlebovirus, Arumowot	AR-1284-64
Phlebovirus, Chagres	JW10
Phlebovirus, Gabek Forest	Sudan 754-61
Phlebovirus, Punta Toro	Balliet
Phlebovirus, Sandfly Fever Sicilian	Strain
Phlebovirus, Toscana	Iss Phl 3

Table 6. Background panel of RNA extracts from mammalian, avian, and arthropod cell lines used in signature screening.

Vero: African Green Monkey Kidney
HeLa:Human epithelial fetal cervical carcinoma
A549: Human carcinoma epithelial lung
SF-9: Insect cells
DF-1: (Chicken embryonic fibroblast)
MRC-5: Human lung fibroblast. P17
RD: rhabdomysarcoma, human muscle
BFK: bovine fetal kidney
MDCK: Madine-Darby canine kidney, epithelial kidney
PK: Porcine kidney, epithelial
C636: Insect cells

Per Signature Limit of Detection for Rift Valley Fever Strains Tested								
		Plaque-forming Units (pfu)*						Picograms total RNA**
	Signature	SA75	Ken 58 (B691)	Maurita-nia OS-1	ZH548	ZH501	SA51	MP-12
Degenerate	SC1	0.10	1.0	0.001	100.0	1.0	0.01	0.10
	SC2	0.10	1.0	0.001	100.0	1.0	0.10	0.10
	M2C2	0.10	1.0	0.001	10.0	0.10	0.01	0.10
	L1C1	0.001	0.10	0.001	100.0	1.0	0.001	0.10
	L1C2	0.01	0.10	0.01	1000.0	1.0	0.001	0.10
	L2C3	1.0	1.0	0.01	1000.0	1.0	1.0	1.0
Non-degenerate	1756318	0.10	1.0	0.10	10.0	0.10	1000.0	0.10
	1756321	1.0	100.0	1.0	100.0	1.0	1000.0	1.0
	1756325	0.10	1.0	0.01	10.0	0.10	1000.0	0.10

Table 7: All tests were conducted on total nucleic acid extracts from clarified viral cultures containing the listed strains. QRT-PCR tests were performed in triplicate on the ABI7500FAST platform as described in materials and methods. An 8-log dilution series of templates ranging from 200pfu/ul to $2X10^{-4}$ pfu/ul (or 20pg/ul to $2X10^{-6}$pg/ul for MP-12) was made for each template. Five microliters of each dilution was spiked into qRT-PCR plates in triplicate. * A plaque-forming unit (PFU) is a measure of the number of particles capable of forming plaques per unit volume, such as virus particles. It is a functional measurement rather than a measurement of the absolute quantity of particles: viral particles that are defective or which fail to infect their target cell will not produce a plaque and thus will not be counted. ** For RVF MP-12, no titer information is provided as multiple attempts to titer this species failed due to lack of consistent CPE. Thus, sensitivity of this assay is reported in pg of total nucleic acid extract from viral culture. No other RNA (such as poly-A RNA used as carrier in extractions) was spiked into samples. For the degenerate signatures, only the primer/probe combinations that were able to detect all strains of RVFV tested are listed.