Genomic surveillance of Rift Valley fever virus

John Juma[1,2], Vagner Fonseca[3,4,5,6], Konongoi Limbaso[1,7], Peter van Heusden[2], Kristina Roesel[1], Rosemary Sang, Bernard Bett[7], Alan Christoffels[2], Tulio de Oliveira[3,4,8,9]and Samuel Oyola[1]

[1]International Livestock Research Institute (ILRI), Nairobi, Kenya
[2]South African National Bioinformatics Institute, South African MRC Bioinformatics Unit. Cape Town, South Africa
[3]KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), School of Laboratory Medicine and Medical Sciences, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa.
[4]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University Stellenbosch, South Africa
[5]Laboratorio de Genética Celular e Molecular, Instituto de Ciências Biologicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil;
[6]Organização Pan-Americana da Saúde/Organização Mundial da Saúde, Brasília, Distrito Federal, Brazil
[7]Centre for Virus Research (CVR), Kenya Medical Research Institute (KEMRI), Nairobi, Kenya
[8]Centre for the AIDS Programme of Research in South Africa (CAPRISA), Durban, South Africa
[9]Department of Global Health, University of Washington, Seattle, WA, USA

BUiLD Boosting Uganda's Investment in Livestock Development

22 September 2022

Introduction

Previous phylogenetic studies on Rift Valley fever virus (RVFV) have identified 15 lineages, designated from A to O, circulating in sub-Saharan Africa.
Genomic surveillance to elucidate genetic diversity of the virus is crucial in understanding the emergence and spread of outbreaks as well as prevention interventions.
We present a computational tool for rapidly classifying and assigning lineages of RVFV isolates using 4 classifiers, namely glycoprotein (gn) and whole genome sequences (L, M and S-segments). The method is available as a web application and a command line tool.

Methods

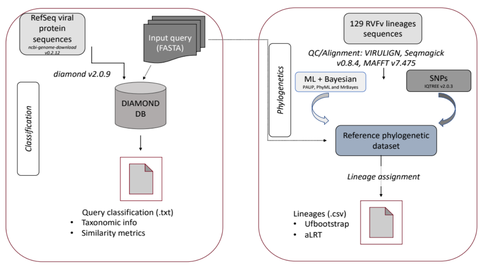

Figure 1. Schematic representation of the command line workflow. The workflow begins with virus classification using DIAMOND and reports the output as a text file with taxonomic information and similarity metrics. Phylogenetic analysis is performed using a default phylogenetic reference dataset generated by Neighbor-Joining (NJ), Maximum likelihood (ML) and Bayesian tree. Users can specify which phylogenetic reference dataset to use. Query sequences are aligned to the reference dataset multiple sequence alignment with MAFFT, and a ML phylogenetic tree is constructed followed by lineage assignment. An output file with the lineage assignment, bootstrap values and likelihood test ratio is generated in comma-separated values (CSV) file format.

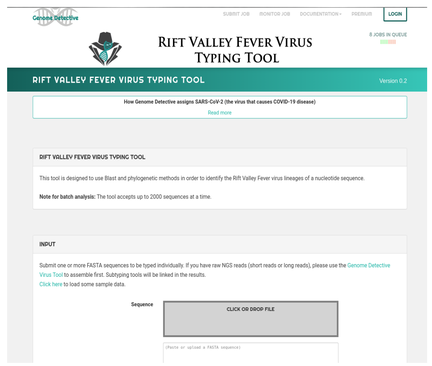

Figure 2. Screenshot of the web interface for RVFV typing tool. The web interface offers a portal for users to perform classification and visualize the results. Users can input upto a maximum of 2000 consensus genome sequences in FASTA format. Results can be queried and visualized interactively, with an option to download the phylogenetic trees in various (.fasta, .nexus and .png) and classification results in comma-separated value format.

```
executor >  local (1)
[f8/b1bfa8] process > RVFVTYPING:FILTER_INPUT (DQ380189)                [100%] 12 of 12, cached: 12 ✔
[eb/94268b] process > RVFVTYPING:GUNZIP_DATABASE (viral.protein.faa.dmnd) [100%] 1 of 1, cached: 1 ✔
[eb/e11c05] process > RVFVTYPING:DIAMOND_BLASTX (MG972973)              [100%] 12 of 12, cached: 12 ✔
[20/064350] process > RVFVTYPING:MAFFT_ALIGN_QUERY (MG972973)           [100%] 12 of 12, cached: 12 ✔
[d4/483b2c] process > RVFVTYPING:IQTREE_QUERY (EU574086.1)              [100%] 12 of 12, cached: 12 ✔
[15/cdb0b1] process > RVFVTYPING:BIOPYTHON_NEWICK_TO_NEXUS (MG972973)   [100%] 12 of 12, cached: 12 ✔
[47/0c1fdb] process > RVFVTYPING:DENDROPY_LINEAGE (MG972973)            [100%] 12 of 12, cached: 12 ✔
[a3/373f8a] process > RVFVTYPING:SNPS_TO_CSV (MG972973)                 [100%] 12 of 12, cached: 12 ✔
[42/393514] process > RVFVTYPING:ORDER_BY_TIPLABELS (MG972973)          [100%] 12 of 12, cached: 12 ✔
[17/6fb4b4] process > RVFVTYPING:PLOT_TREE_SNPS (MG972973)              [100%] 12 of 12, cached: 12 ✔
[06/9c6024] process > RVFVTYPING:PLOT_TREE_MSA (MG972973)               [100%] 12 of 12, cached: 12 ✔
[2e/7f0bdc] process > RVFVTYPING:REPORT_WITH_LINEAGE (report)           [100%] 1 of 1, cached: 1 ✔
[d1/ad4c59] process > RVFVTYPING:GET_SOFTWARE_VERSIONS                  [100%] 1 of 1 ✔
-[rvfvtyping] Pipeline completed successfully-
```

```
Query,Lineage,aLRT,UFbootstrap,Length,Ns(%),Note,Year_first,Year_last,Countries
HM587045,B,86,97,490,0.00,assigned (bootstrap value >= 70),1972,1972,Kenya
HM587125,O,93,97,490,0.00,assigned (bootstrap value >= 70),1951,1951,South Africa
AF134495,N,88,86,738,0.00,assigned (bootstrap value >= 70),1975,1993,Burkina Faso;Senegal;Mauritania
DQ380222,J,76,38,3885,0.00,unassigned (bootstrap value < 70),,,
HM587108,I,87,92,490,0.00,assigned (bootstrap value >= 70),1955,1956,South Africa
HM587118,L,99,100,490,0.00,assigned (bootstrap value >= 70),1963,1995,Zimbabwe;Egypt;South Africa;Kenya
DQ380221,D,92,98,3885,0.00,assigned (bootstrap value >= 70),1973,1973,CAR
```

Figure 3. Screenshot of the command line tool for RVFV typing tool. The command line tool offer a versatile ability to query and assign lineages to multiple sequences using any of the 4 classifiers. It runs a series of steps/processes in a asynchronous manner to allow for parallelization. The output of the command line tool include a text file showing the assigned lineage with associated phylogenetic statistical measures. The tool also provides visual representations of the tree in .pdf file format as well as alignments in fasta and nexus formats.

Findings

Congruency in tree topology of maximum likelihood phylogenetic trees indicating low occurrence of genetic recombination/reassortment.

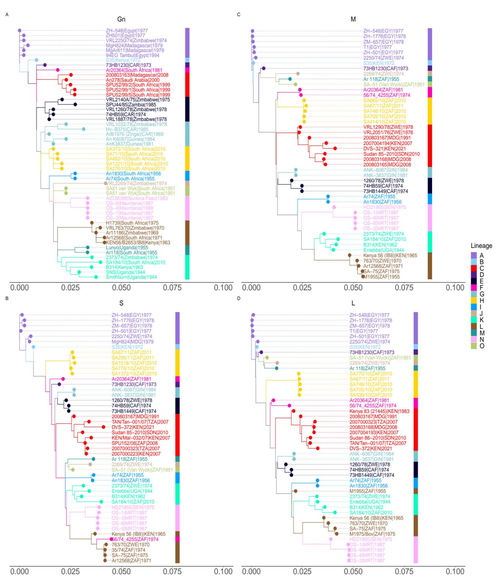

Figure 4. Phylogenetic analysis using Gn and whole genome (L, M & S) segment classifiers. (A-D) Maximum likelihood (ML) phylogenetic trees inferred from the representative sequences for all lineages within the (A) 51 sequences of the glycoprotein (490 bp) gene aligned with MAFFT and ML tree inferred under the GTR+Γ+G substitution model, (B) 47 sequences of the Small (S) segment (1690 bp), (C) 47 sequences of the Medium (M) segment (3885 bp) and (D) 47 sequences of the Large (L) segment (6404 bp). All the trees show similar topology for all the lineages.

High sensitivity, specificity and accuracy in assignment of lineages using the glycoprotein Gn classifier.

Lineage	*Known*	*TP*	*TN*	*FP*	*FN*	*TPR*	*FPR*	*ACC*
A	13	13	115	0	0	100.0	0.0	100.0
B	1	1	127	0	0	100.0	0.0	100.0
C	44	44	84	0	0	100.0	0.0	100.0
D	1	1	127	0	0	100.0	0.0	100.0
E	7	7	121	0	0	100.0	0.0	100.0
F	1	1	127	0	0	100.0	0.0	100.0
G	8	8	120	0	0	100.0	0.0	100.0
H	12	12	116	0	0	100.0	0.0	100.0
I	2	2	126	0	0	100.0	0.0	100.0
J	1	1	127	0	0	100.0	0.0	100.0
K	11	11	117	0	0	100.0	0.0	100.0
L	10	10	118	0	0	100.0	0.0	100.0
M	2	2	126	0	0	100.0	0.0	100.0
N	13	13	115	0	0	100.0	0.0	100.0
O	2	2	126	0	0	100.0	0.0	100.0

Table 1. Validation/testing of the RVFV Typing tool to classify partial and whole genome sequences (*n*=128) using glycoprotein sequences. The classification results were compared to manual phylogenetic analysis. Abbreviations as used in this table: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, TPR = True Positive Rate, FPR = False Positive Rate, ACC=Accuracy.

Lineage assignment in outbreak clinical samples using the 4 classifiers (L, M and S whole genome segments and Gn) indicate correct classification.

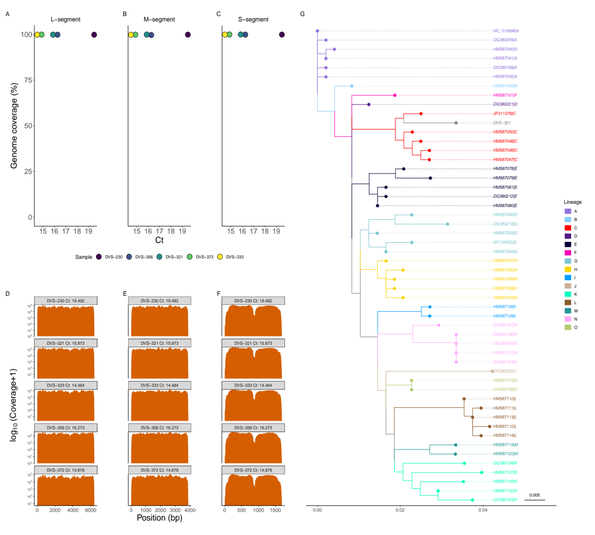

Figure 5. Genome sequencing and phylogenetic analysis. (A-C) RT-qPCR cycle threshold (Ct) values (*x*-axis) plotted against percent genome covered (*y*-axis) for L, M and S-segments of the RVFV genome. **(D-F), Genome** coverage (log10 transformed) along genomic positions in base pairs. **(G)** Maximum likelihood phylogenetic tree indicating the different clades corresponding to the fifteen major lineages and showing where a query sequence (DVS-321) is clustered in the tree.

Conclusions and limitations

RVFV Typing tool presented here allows for fast and accurate classification of RVFV species and lineages within a few minutes using the whole genome (L, M, & S-segments) and/or the partial glycoprotein Gn (490 bp) sequences.

Contribution to Uganda's livestock development agenda

- Capacity building in next generation sequencing of pathogens and field epidemiological studies.
- Timely determination of circulating lineages.

Corresponding author: John Juma
J.Juma@cgiar.org
ILRI
P.O. Box 30709, Nairobi, Kenya
+254 711 033 860

ILRI thanks all donors and organizations which globally support its work through their contributions to the CGIAR Trust Fund.

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE
Federal Ministry for Economic Cooperation and Development BMZ
CGIAR RESEARCH PROGRAM ON Agriculture for Nutrition and Health
CGIAR RESEARCH PROGRAM ON Livestock
UNIVERSITY of the WESTERN CAPE