Making Genomic Surveillance Deliver: A Lineage Classification and Nomenclature System to Inform Rabies Elimination

Campbell, K.[1], Gifford, R.J.[3], Singer, J.[3], Hill, V.[2], O'Toole, A.[2], Rambaut, A.[2], Hampson, K.[1] & Brunker, K.[1]

[1]Institute of Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, G12 8QQ, UK

[2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3FL, UK

[3]MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow G61 1QH, UK

Corresponding author information: k.campbell.1@research.gla.ac.uk

Abstract

The availability of pathogen sequence data and use of genomic surveillance is rapidly increasing. Genomic tools and classification systems need updating to reflect this. Here, rabies virus is used as an example to showcase the potential value of updated genomic tools to enhance surveillance to better understand epidemiological dynamics and improve disease control. Previous studies have described the evolutionary history of rabies virus; however, the resulting taxonomy lacks the definition necessary to identify incursions, lineage turnover and transmission routes at high resolution. Here we propose a lineage classification system based on the dynamic nomenclature used for SARS-CoV-2, defining a lineage by phylogenetic methods, for tracking virus spread and comparing sequences across geographic areas. We demonstrate this system through application to the globally distributed Cosmopolitan clade of rabies virus, defining 73 total lineages within the clade, beyond the 22 previously reported. We further show how integration of this tool with a new rabies virus sequence data resource (RABV-GLUE) enables rapid application, for example, highlighting lineage dynamics relevant to control and elimination programmes, such as identifying importations and their sources, and areas of persistence and transmission, including transboundary incursions. This system and the tools developed should be useful for coordinating and targeting control programmes and monitoring progress as we work towards eliminating dog-mediated rabies, as well as having potential for broad application to the surveillance of other viruses.

Keywords: Surveillance; Zero by 30; Lyssavirus; Sequencing; Canine rabies; Rabies control

Author Summary

The importance of the ability to track the diversity and spread of viruses in a universal way that can be clearly communicated has been highlighted during the SARS-CoV-2 pandemic. This, accompanied with the increase in the availability and use of pathogen sequence data, means the development of new genomic tools and classification systems can strengthen outbreak response and disease control. Here, we present an easy-to-use objective and transferable classification tool for tracking viruses at high resolution. We use rabies virus, a neglected zoonotic disease that causes around 59,000 human deaths each year, as an example use case of this tool. Applying our tool to a global clade of rabies virus, we find an over 200% increase in the definition at which we can classify the virus, allowing us to identify areas of persistence and transmission that were not previously apparent, and patterns of virus spread. Insights from the application of this tool should prove valuable in targeting vaccination campaigns and improving surveillance as countries work towards the elimination of dog-mediated rabies.

Introduction

Rabies virus (RABV) causes around 59,000 deaths [1] and costs in excess of $8.6 billion per year [2], with a near 100% mortality rate after the onset of symptoms [3]. Post-exposure prophylaxis is highly effective in preventing rabies if administered quickly following exposure [4] but this is not always possible given its high cost and limited accessibility. Rabies can occur in all species of mammal, but up to 99% of human rabies cases arise from bites from infected domestic dogs [5,6]. Vaccinating dogs to interrupt transmission is therefore paramount [7] and a major focus of 'Zero by 30', the global strategy to eliminate human deaths from dog-mediated rabies by 2030 [8]. The focus of 'Zero by 30' is on dog-mediated rabies, but spill over from dogs into other carnivores often occurs, generally causing only short-lived chains of transmission [9]. Genomic surveillance is therefore vital to identify and monitor any wildlife reservoirs that may emerge [10].

To achieve the 'Zero by 30' goal, effective and coordinated surveillance is essential. Genomic surveillance can complement routine epidemiological surveillance through the insights it can provide on the lineages circulating in an area and any sources of incursions [11]. Sequence data proved useful in understanding rabies dynamics in Bangui, the capital of the Central African Republic [12]. Instead of sustained transmission in the city, local extinction occurred on three occasions with introductions from surrounding areas reseeding circulation, showing the need to expand control efforts across a larger geographic area [12]. Genomic surveillance can also reveal host shifts and other unusual dynamics. When rabies in Arctic foxes underwent a host shift to infect red foxes (*Vulpes vulpes*), the virus spread rapidly throughout Canada, resulting in epidemics in the 1950s and 1960s before being apparently eliminated in 1990 [13]. However, between 2015 and 2017 a number of rabies cases occurred in wildlife which were genetically similar to sequences from these historic fox rabies epidemics [13]. Analysis revealed that although several lineages were eliminated, one persisted and underwent a host switch to skunks [13].

To monitor how an epidemiological situation is changing first requires characterization of the current situation. Several well-defined RABV clades circulate globally, within two major phylogenetic

groups; bat-related and dog-related [1]. The dog-related group is split into between 5 and 7 different clades, depending upon the classification [1,14]. This includes the Cosmopolitan clade; over 9500 publicly available sequences (including sequences of all lengths) split into 22 subclades, present in over 100 countries [14,16]. These discrepancies in clade numbers are illustrative of issues surrounding the interpretation and classification of RABV phylogenetic data as a universal naming system does not extend past high-level classification, with no set rules for defining lineages. Additionally, genomic data availability varies. Increasingly studies are focusing on whole genome sequences (WGS) given the greater resolution they provide, but the vast majority of studies thus far have generated shorter, partial gene sequences [15–17]. A lineage designation system therefore needs to be able to incorporate all of this data to provide maximal contextual information. These terms, and others relevant to this study, are defined in box 1.

Box 1 - Definitions

Clade – Monophyletic group of sequences with a single ancestor.

Subclade – A smaller monophyletic group contained within a larger clade.

Lineage – A group of genetically related sequences defined by statistical support of their placement in a phylogeny and genetic differences from a common ancestor.

> **Major lineage** – A lineage named with a letter - e.g. A1 or Cosmopolitan AF1b_A1, that can be the first iteration of a lineage, or a lineage that has undergone significant evolution to become a new major lineage.
>
> **Minor lineage** – A lineage named with numbers (following the major clade nomenclature) - e.g. A1.1.1 or Cosmopolitan AF1b_A1.1.1 This is a major lineage that has undergone evolution, but not enough to become a new major lineage.

Lineage designation - An initial step to designate lineages based on a set of reference sequences, or the first set of data from an area, defined by an existing set of rules, and to name them accordingly. This is completed once to form a reference set of sequences to be used for lineage assignment, and may need to be updated as genetic diversity accumulates.

Lineage assignment - Identifying which existing lineage, defined by the initial lineage designation step, a new sequence belongs to.

MAD DOG - Method for Assignment, Definition and Designation Of Global rabies virus lineages; the method and corresponding tools presented in this study.

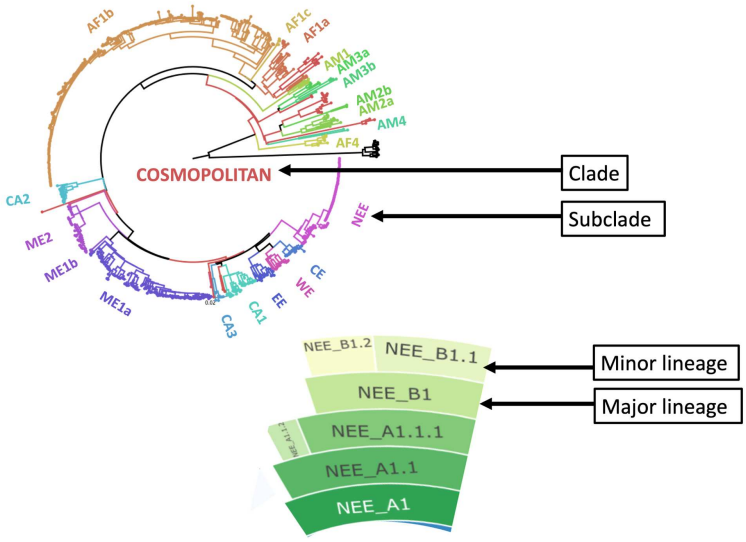


Figure 1. Illustration of rabies virus clades, subclades and lineages. *Maximum likelihood tree of all publically available whole genome rabies virus sequences (n=650) coloured by previously identified subclades [14]. The NEE subclade has been expanded to show major and minor lineages from the updated MADDOG classification system.*

Although the current phylogenetic classifications generally work well at a global level, they lack resolution for surveillance at more local or regional scales. The existing system generally represents where the different clades originally emerged and their early geographical distribution [1]. However, these clades do not always remain isolated to particular areas, and their dispersal is affected by human movements [18], leading to frequent introductions of lineages to new areas and their subsequent co-circulation [18]. Moreover, control and elimination efforts will further affect the distribution and diversity of circulating lineages. In many regions, a limited number of subclades appear to circulate [19], therefore, without higher lineage resolution it becomes difficult to identify patterns like lineage extinction that are to be expected, and therefore to monitor the impacts of control.

Rambaut et al. (2020) propose a universal virus nomenclature system to address these issues, which they apply to SARS-CoV-2 [20]. The system outlines a set of rules for classifying and naming viral lineages to produce a standardised tool to identify viral diversity on global and local scales, to track lineage emergence and transmission, and to allow for coherent updates as lineage turnover occurs [20]. Here, we adapt and apply these rules to a large diverse major clade of RABV to produce an updated, dynamic, classification system for the virus. The increased definition provided allows for more detailed genomic surveillance that can be used to monitor the circulation of viral lineages and identify incursions, unusual transmission routes and potential host shift events.

Box 2 - RABV-GLUE

RABV-GLUE is a 'sequence data resource' developed to support rabies elimination efforts by enabling efficient dissemination and utilization of RABV sequence data using GLUE - a bioinformatics environment for managing and interpreting virus sequence data [21, 22]. GLUE supports development of virus-specific 'projects' comprising curated sets of sequences, genome feature annotations, alignments, reference clades and phylogenies with associated metadata, with options to upload sequences to GenBank. Loading projects into the GLUE 'engine' creates a relational database representing complex semantic links between data items so computational analyses can be precisely and widely replicated.

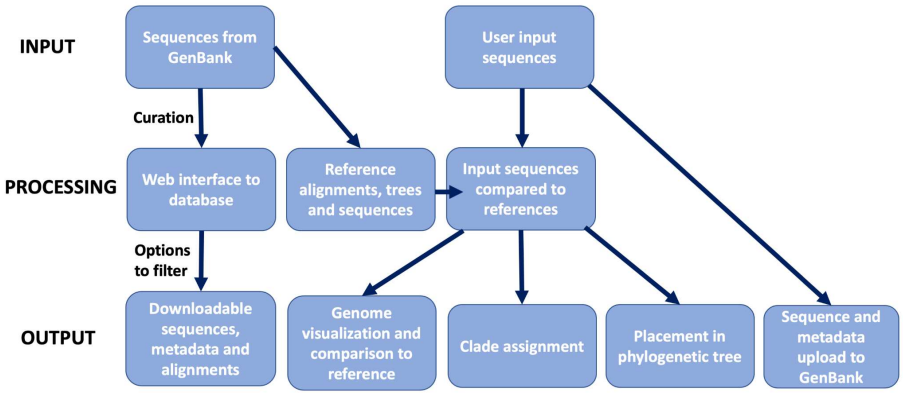


Figure 2. Schematic of the online capabilities of RABV-GLUE. *RABV-GLUE can be used to obtain sequences from GenBank and has options to filter and download sequences of interest, and for phylogenetic interpretation of user input sequences.*

RABV-GLUE projects can be installed locally on all commonly used computing platforms and are fully containerised via Docker [23]. GLUE's command layer provides a mechanism for retrieving and analysing data by coordinating interactions between the RABV-GLUE database and commonly used bioinformatics software tools (e.g. MAAFT, RAXML). This allows experienced bioinformaticians to quickly establish local RABV sequence databases and integrate these resources into existing bioinformatic pipelines, tailoring functionalities to their specific needs. Hosting the RABV-GLUE project in an openly accessible online version control system (e.g., GitHub) provides a mechanism for managing its ongoing development by multiple collaborators, following practices established in the software industry. GLUE projects can also be developed into interactive, user-friendly web services through a graphical user interface. An online interface to RABV-GLUE is available at http://rabv-glue.cvr.gla.ac.uk/ (Fig 2). Here we use sequence data and classifications from RABV-GLUE, to provide the basis for the development of an updated classification system which will be integrated into RABV-GLUE, enhancing its capacity to provide detailed, high resolution lineage information about user input sequences.

Methods

Data Collection and Processing

All available RABV sequences (n=23,386), irrespective of sequence length and clade, and their associated metadata were downloaded from the RABV-GLUE website (http://rabv.glue.cvr.ac.uk). Many metadata entries stripped from Genbank and added to the RABV-GLUE database were missing information necessary for analysis, including the sample collection year, host, and location (n=9650). Records with missing information were searched manually and any information that could be found from primary publications was updated into RABV-GLUE. The dataset of all sequences was then filtered to only include sequences designated to the Cosmopolitan clade by RABV-GLUE, excluding vaccine strains. For the purposes of these analyses we considered WGS, covering >90% of the genome (at least 10,000 nucleotides (nt)), and nucleoprotein (N) gene sequences (1300 nt), thus excluding smaller partial gene fragments.

Lineage Designation

Sequences were aligned using MAFFT [24] with the FFT-NS-2 algorithm [24]. Maximum Likelihood trees were constructed using IQTREE2 with model selection [25] and 100 bootstrap replicates [26]. Ancestral sequence reconstruction was then undertaken using Treetime ancestral [27].

A custom R function for lineage designation was developed which requires the tree, corresponding alignment, ancestral sequences and metadata. This returns summary statistics about each sequence, its lineage designation, and details about each of the designated lineages. Lineage defining nodes are identified according to thresholds on bootstrap support (>70) and cluster size (>10 descendents of >95% coverage at genome level or of the gene of interest, excluding gaps and ambiguous bases). For each node still in consideration, the ancestral sequence is extracted. To define a new lineage, there must be at least one common mutation between all descendents that is different to the ancestral sequence. The algorithm for lineage designation is summarised in Fig 3. The parameters for designation of partial genome sequences were refined to optimise the comparability with the whole genome designations.

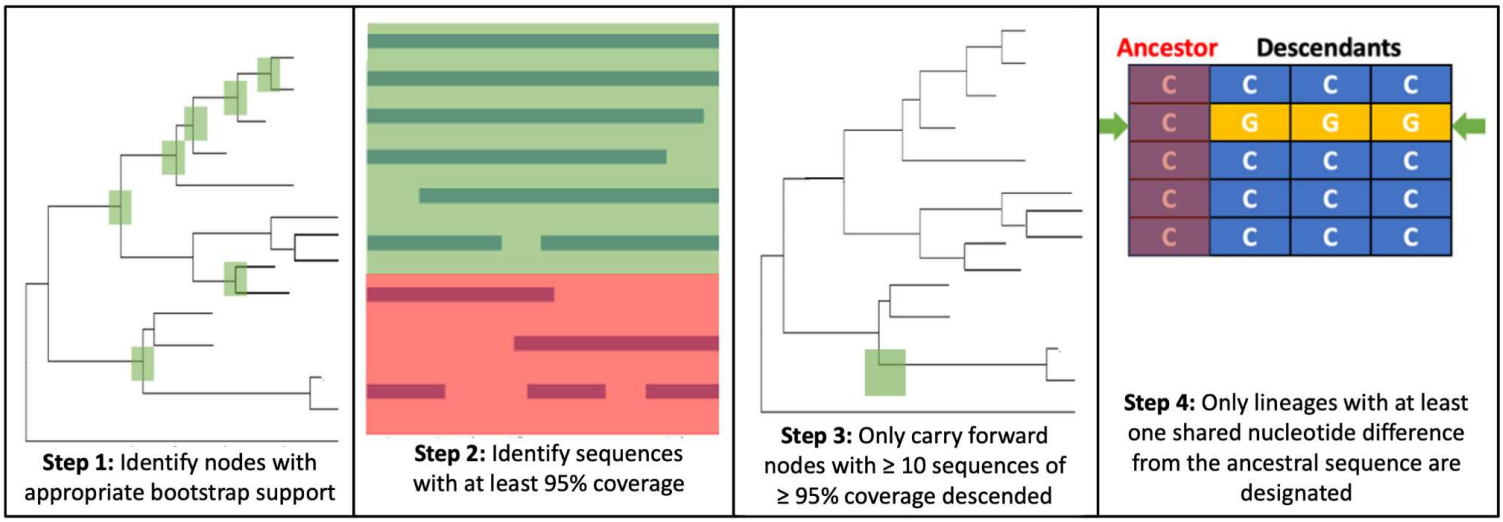


Figure 3. Summary of lineage designation steps.

Existing phylogenetic groupings are already in place for RABV [14,28] and sequences are automatically defined by these clades in RABV-GLUE which acts as a useful baseline for further classification , and subsequent MAD DOG lineage designations are named according to the rules in Rambaut et al. [21], starting from lineage A1 (which in this instance would be Cosmopolitan A1). Any lineages descended from this become A1.1 or A1.2 etc and any descended from these become A1.1.1, A1.2.1 etc. After 3 iterations a lineage becomes a new major lineage – A1.1.1.1 = B1.

Lineage Assignment

Once the designation step has been performed to construct a set of reference sequences with defined lineages, new sequences can be compared to the reference set and assigned to lineages. New sequences (imported individually or in bulk) are added to the existing reference alignment using MAFFT and for each new sequence the two closest references are identified and used to transfer a lineage assignment.

MAD DOG

MAD DOG has been developed as a command-line based tool to undertake all the steps for lineage designation and assignment outlined above, given a set of input sequences (and corresponding metadata for designation).The assignment tool also outputs information about the assigned lineage including which countries it has been sampled in and when the first and most recent sequences for that lineage were recorded. Additionally, this can be implemented through an R package (MADDOG), full details of which can be found at DOI: 10.5281/zenodo.5503916.

MAD DOG was used to designate lineages for sequence subsets from the Cosmopolitan clade, with its performance tested on whole and partial genome sequences. The resulting designations were explored on a global scale and on a specific geographic area (Tanzania) to test functionality and interpret lineage assignments.

Results

Cosmopolitan Dataset

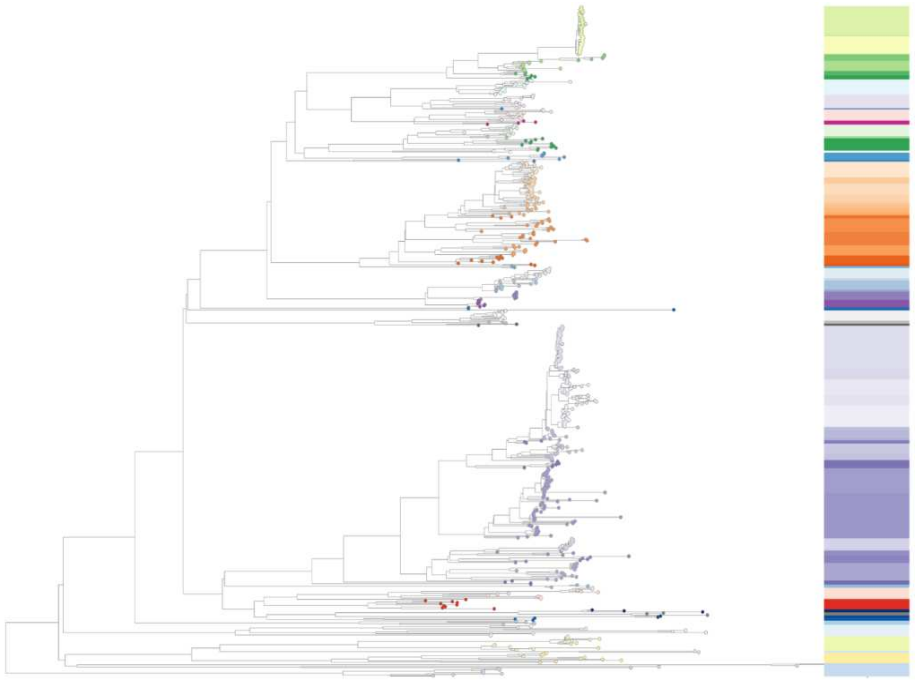
The Cosmopolitan dataset comprised 650 WGS, excluding vaccine strains, spanning 46 countries from 1950-2018. The N gene data comprised an additional 1476 sequences spanning an additional 25 countries from 1950-2020. Running the lineage designation on Cosmopolitan WGS resulted in 73 lineages, more than a 3-fold increase in definition compared to 22 previously defined subclades (Fig 4). The baseline MAD DOG designation agreed with the previously determined global phylogenetic groupings, with the exception of clusters containing <10 sequences (a required threshold for MAD DOG lineage designation). However, our tool provided a much deeper characterisation beyond the subclade level [1]. Some subclades are split into multiple lineages providing increased definition; four showed at least a 5-fold increase in definition, and two showed over a 10-fold increase, in line with sampling density (Table 1, Fig 4b). The AF1b subclade (purple in Figs 4 and 5) is split into 20 lineages, and the ME1a (orange) subclade is split into 13, indicating more than a 1000% increase in definition in both cases. Some subclades show no further definition due to the small number of sequences available to represent this subset of RABV diversity.

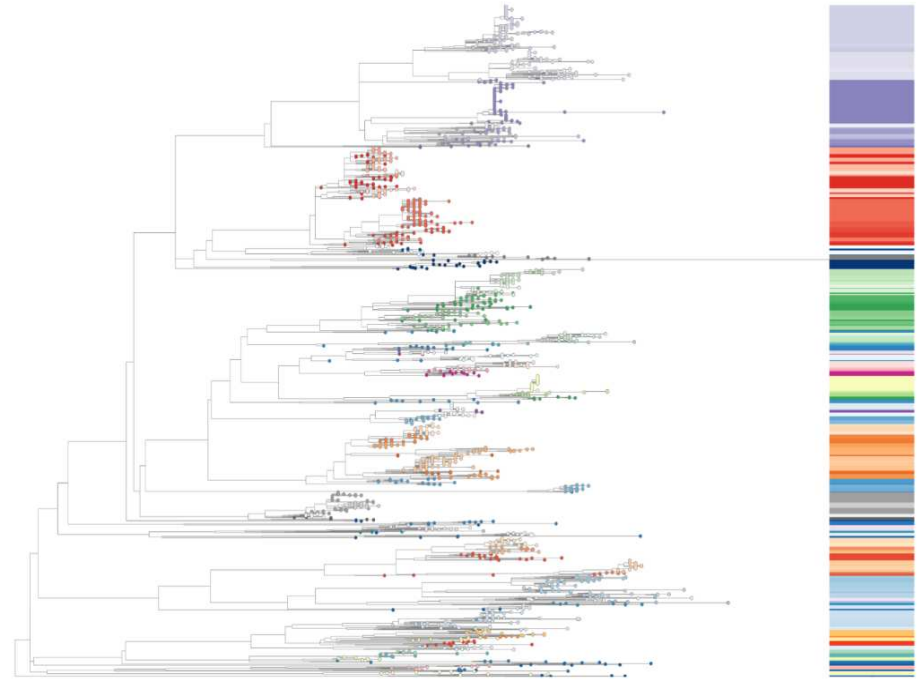
	Whole genome			N gene		
Subclade	**Lineages**	**Number of sequences**	**Year of first sequence**	**Lineages**	**Number of sequences**	**Year of first sequence**
AF1a	2	24	1984	14	377	1982
AF1b	20	251	1981	14	471	1981
AF1c	0	3	1986	0	6	1986
AF4	0	8	1950	1	13	1950
AM1	0	4	1982	3	58	1981
AM2a	1	10	1991	4	136	1990
AM2b	0	2	2009	0	24	1986
AM3a	0	5	1986	11	172	1985
AM3b	0	4	1986	3	108	1986
AM4	0	4	1974	1	19	1974
CA1	3	25	1974	14	224	1974
CA2	3	15	1993	4	106	1993
CA3	0	7	1991	5	45	1991
CE	1	15	1990	1	19	1985
EE	9	86	1986	10	155	1977
ME1a	13	101	1976	15	195	1976
ME1b	0	2	1993	0	7	1993
ME2	6	38	1989	3	66	1989
NEE	7	71	1986	5	78	1985
WE	1	14	1986	2	148	1974
YUGCOW	0	2	1978	0	6	1978
YUGFOX	0	1	1972	0	12	1972

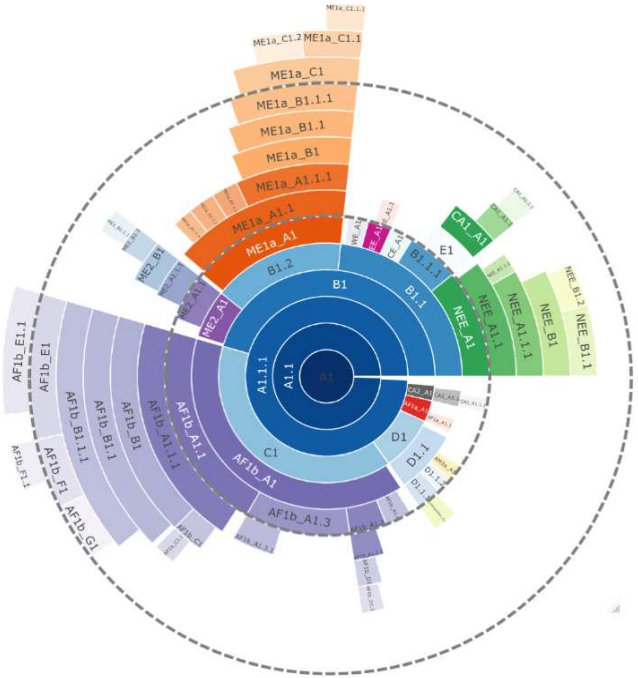
Table 1. Details of numbers of lineages and sequences available for each Cosmopolitan subclade at whole genome and N gene level.

Lineage designation was performed on an N gene subset of the same data (the 650 sequences used for WGS designation) to allow a direct comparison between lineage designations at different levels of sequence data resolution. This resulted in 55 MAD DOG designations (Figure S2), a 25% overall decrease in definition compared to using WGS, highlighting that the sequence length impacts the designations. The loss of definition applies to some subclades more than others. Various numbers of minimum sequences and minimum bootstrap support values to define a lineage were tested on the N gene subset and the WGS designations to optimise the comparability of the two. Once the parameters for designation at N gene level were refined at minimum 10 sequences and bootstrap support of 70 using this subset, it was applied to all available Cosmopolitan N gene sequences; an increase from 650 to 2126 sequences. This resulted in an additional 75 designations above the WGS designations due to the increased volume of sequences (Fig 4 and 5, Table 1), highlighting how data volume and coverage also affects the designations.

The geographic distribution of lineages can be seen in Fig 5, with full details in Table S1. It appears that Europe and Southern Africa have good coverage, especially at whole genome level. However, only the Cosmopolitan clade is included here and therefore the overall global RABV sequencing coverage cannot be assessed from this alone.







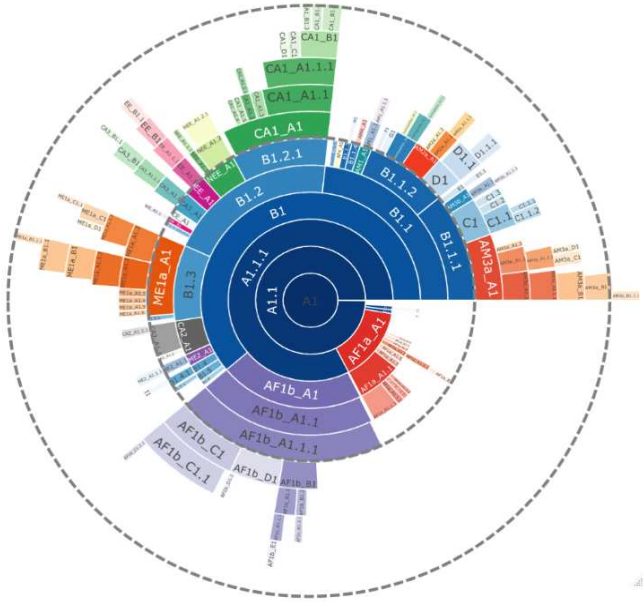


Figure 4. Lineage designations of WGS and N gene sequences from the RABV Cosmopolitan clade. *Maximum likelihood tree showing the lineage positions, rooted using an outgroup of 10 Asian SEA1b sequences, and hierarchical relationships with dashed lines indicating 5 lineage iterations of WGS (n=650)* ***(above)*** *and N gene sequences (n=2126)* ***(below)****. Sequences obtained from RABV-GLUE.*

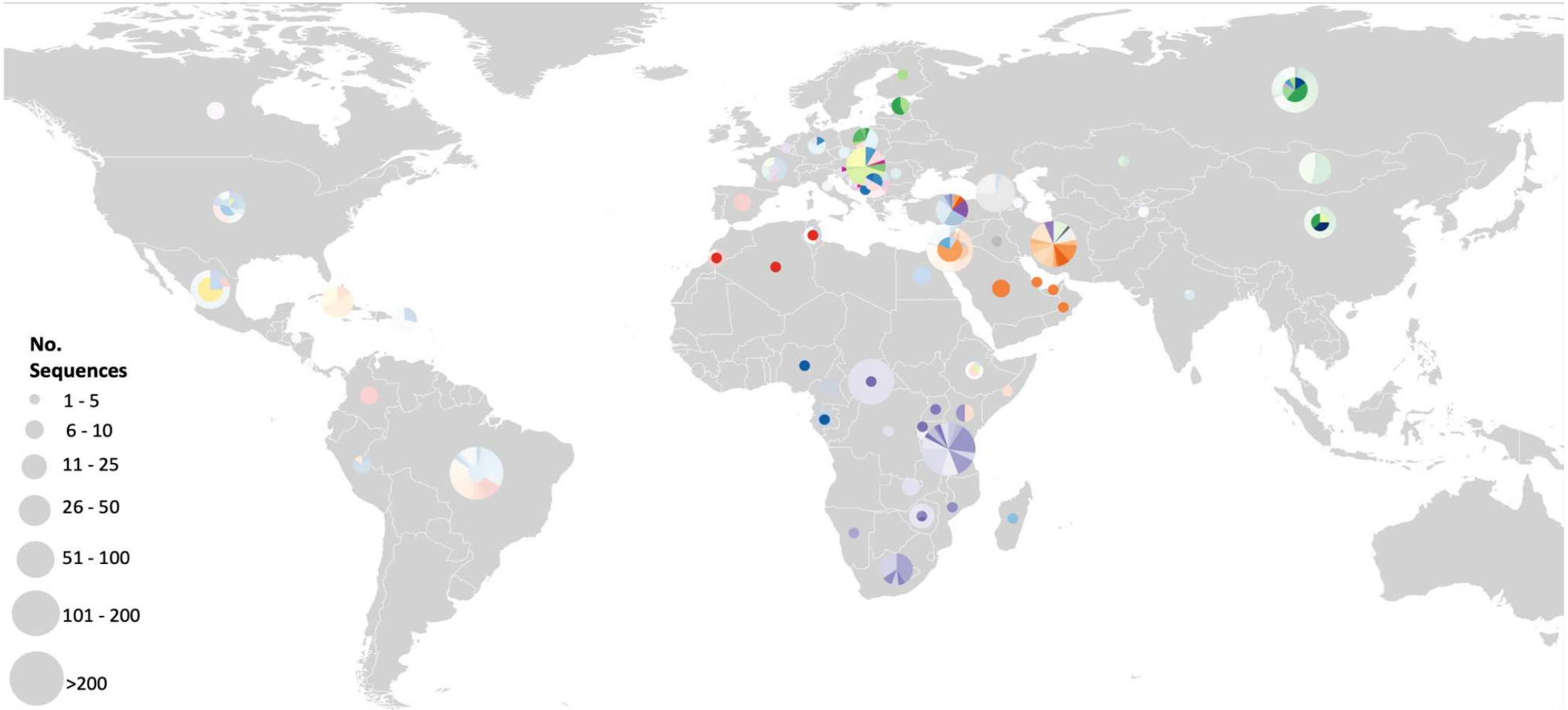


Figure 5. Distribution of WGS and N gene RABV lineages from the Cosmopolitan clade. *More vibrant colours indicate WGS, with transparent overlays representing N gene sequences. Circle size indicates the number of sequences, with circles plotted to the centroid of each country. World map cropped to only show areas with Cosmopolitan sequences.*

WGS capture greater variation, and are able to differentiate between samples that are identical at a partial genome level [16]. In this study, although all sequences were unique at the whole genome level, only 78% of these were unique at an N gene level (n=509). While whole genomes offer better phylogenetic resolution, the N gene is commonly targeted by diagnostic laboratories providing a greater number of sequences and wider spatio-temporal coverage to detect and define lineages not captured by the still limited number of WGS. In total, there are 1469 WGS from 82 countries available on RABV-GLUE but 7304 N gene sequences from 109 countries (Figure S2). This highlights the volume of additional data available when using partial genome sequences. However, the proportion of studies sequencing RABV whole genomes is increasing, as are studies using at least N gene length sequences instead of smaller fragments (Figure S2).

Example 1 - Identifying the origin of cases

The higher resolution of this classification system can be used to 'zoom in' on unusual cases and look at the historical and geographical context of a lineage. Here, we illustrate the example of a human rabies case (GenBank accession: KC737850) reported from the USA in 2011 [28], where dog-associated rabies has been eliminated but wildlife variants are endemic.

Figure 6. Detection of lineage AM3a_A1.3 from 1986 to 2018 highlighting an imported human rabies case. *Bright red = information from WGS lineage designation, darker red = additional information from N gene lineage designation. Dated points represent first records of the lineage in the country, with arrows suggesting the likely source of introductions. The star indicates the human rabies case.*

The case report explains that this individual moved from Brazil to the USA, and many years later developed rabies [28]. In this situation the most likely cause of rabies was either exposure to wildlife rabies in the USA, or exposure to rabies in Brazil, which seemed unlikely given the long incubation period. Sequence data shows that the latter scenario is correct (Fig 6), with the RABV sequence from the patient showing high similarity to dog RABV sequences in Brazil as reported by Boland et al. [29]. Using additional N gene sequences, we can examine this lineage in more detail. A sequence from Peru in 1999 (accession: KU938904) (Fig 6), was listed in RABV-GLUE as being from the common vampire bat, *Desmodus rotundus* but the original GenBank record states 'livestock case infected by bat'. In this case it appears to have been misassigned as a bat host by GLUE's automatic curation from Genbank. However, the identification of a dog lineage in a bat is unusual. Follow up with researchers from the original study revealed this was mislabelled on Genbank and was in fact most likely a livestock case infected by a dog [29]. The lineage classification system allows resolution of these unusual cases, and identification of sequences that are incorrectly labelled.

Example 2 - Reconstructing historical RABV spread

Rabies was first documented in Tanzania in the 1930's [30]**.** Although present in North Africa for centuries, the virus is thought to have become endemic and spread within sub-Saharan Africa following European colonisation in the twentieth century [30]. The literature suggests at least two historical introductions to Tanzania; from the North and the South of the country [18]. With the classification system, it is possible to identify spread over time on a finer scale, and therefore evaluate potential support for these theories.

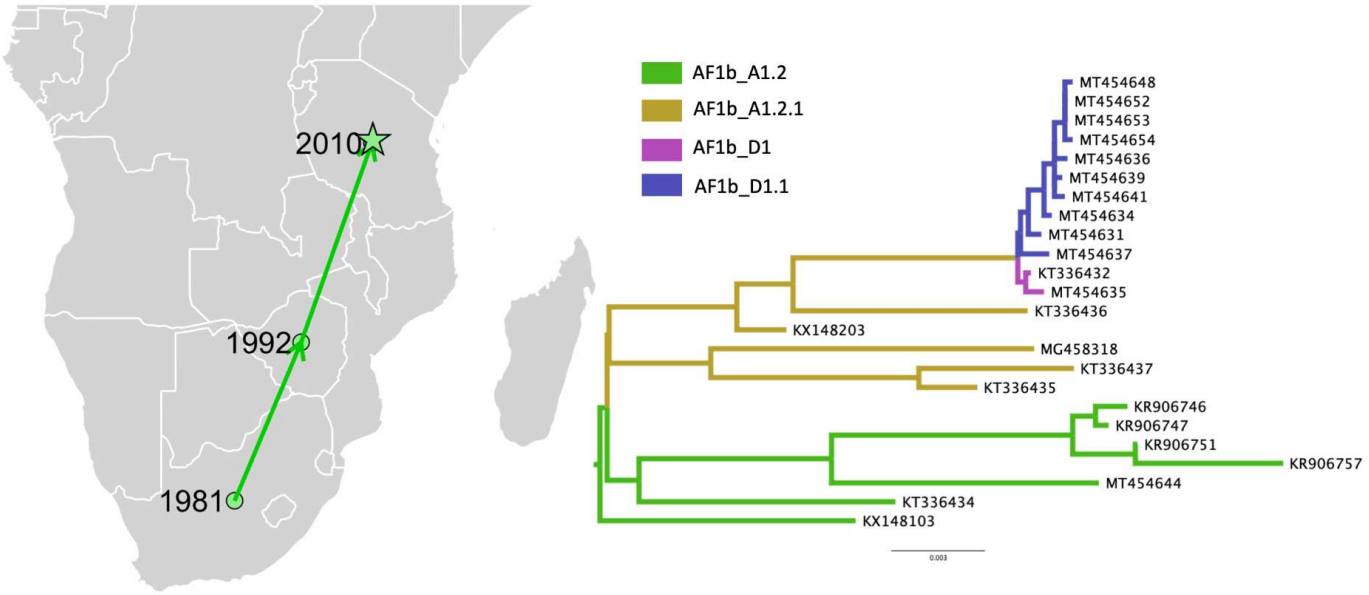


Figure 7. Inferred spread of lineage Cosmopolitan AF1b_A1.2. Left: *Countries where lineage AF1b_A1.2 was recovered from over time (n. seq=7). Dated points represent first records of the lineage in the country, with arrows suggesting the likely route of spread. Star indicates case of interest.* ***Right:*** *Subtree of AF1b A1.2 lineage and descendants (lineages AF1b A1.2.1, AF1b D1 and AF1b D1.1)*

The lineage AF1b_A1.2 appears to support the spread of RABV into Tanzania from southern Africa (Fig 7a), with the first sequence from South Africa in 1981 (KX148103). By 1992 the lineage was detected in Zimbabwe (KT336434), as part of a study looking at potential incursions between South Africa and Zimbabwe [31], before being sequenced in Tanzania in 2010 [20]. Routine monitoring also allows us to identify that the lineage has persisted in South Africa, as the sequence MT454644 from 2017 is designated to the AF1b_A1.2 lineage [32].

Local Lineage Designation Case Study - Tanzania

The Tanzania dataset comprised 205 WGS from 1996-2018. Running the lineage designation on these sequences alone i.e. without wider geographic context, limited the designation of less commonly sampled lineages that may be more prevalent in other countries. The inclusion of sequences from Eastern and Southern Africa (known to belong to the same global subclade) resulted in an additional 2 lineage designations that could not be defined using Tanzania sequences alone. Therefore, additional relevant context to provide an informative reference set for initial lineage designation proved to be important for consistent, higher resolution lineage designations. This revealed 16 lineages present in Tanzania (Table S5) from the Cosmopolitan AF1b subclade, in contrast to only 14 defined with Tanzania data alone. Six lineages are specific to the Serengeti District, 1 to the Pwani Region and the remaining 7 are present in multiple areas (Fig 8). Of the 16 lineages, 11 have been seen in the Serengeti District. In 2010, there were 10 lineages circulating that had been detected in the Serengeti District. By 2017, only 3 were still evident.

Of the sequences from Tanzania, 203 are unique at the whole genome level, but only 117 are unique at N gene level. The increased use of whole genome sequencing is pronounced in Tanzania (Figure S7); most Tanzanian sequences available on RABV-GLUE for the last decade are whole genome, with all sequences since 2005 being at least N gene. There are only an additional 21 N gene sequences available, therefore using N genes sequences provides little additional information, and results in loss of definition over WGS.

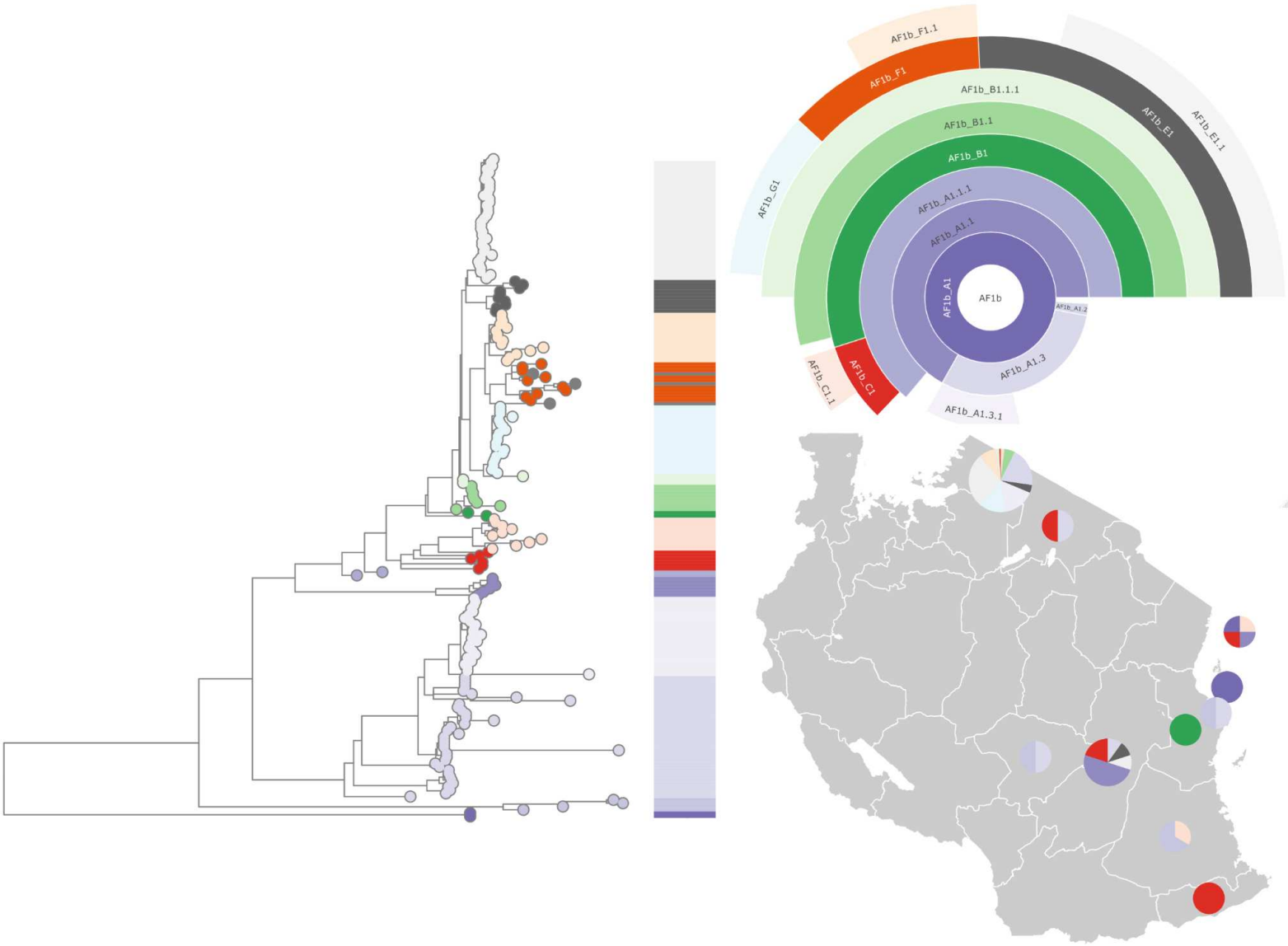


Figure 8. Lineage designations of 205 rabies virus genomes from Tanzania. *Maximum likelihood tree, hierarchical relationships, and distribution of lineages. Sequences obtained from RABV-GLUE.*

Discussion

We aimed to apply the dynamic nomenclature system proposed by Rambaut et al. (2020) to RABV and to investigate how this classification system could be useful in surveillance applications, focusing on supporting the 'Zero by 30' goal. Implementing this classification system allows for much greater resolution of circulating RABV lineages at both the whole genome and N gene levels. Whole genome level designations and assignments provide most resolution (>200% overall increase, with over 20-fold in some areas) and accurate assignments. However, unlike SARS-CoV-2, which is a new pathogen with limited accumulated diversity over a recent time frame, rabies virus is an endemic pathogen that has circulated in some areas for decades. Therefore, our analysis highlighted the importance of considering partial genome data (N gene sequences) to incorporate the considerable historical data available. The case studies and examples presented here illustrate how higher resolution lineage designations enable greater epidemiological interpretation, for example, in identifying the origins of cases and for surveillance to monitor lineage introductions and extinctions as elimination is approached.

There are challenges in adapting the universal classification system developed for SARS-CoV-2 to RABV [20]. The spatial and temporal density of sequences is very different. More than 4 million WGS from over 185 countries are available for SARS-CoV-2 in a time span of less than two years [34,35], whereas RABV sequences span 7 decades but are much sparser. Moreover, SARS-CoV-2 lineages were designated from the emergence of this pathogen in human populations and have been updated as variation accumulated. In contrast the RABV lineage designation is retrospective; accounting for greater diversity and time. As the naming system calls for a new major lineage after 4 iterations of minor lineages (A1.1.1.1 = B1), the names do not fully reflect the depth of diversity and evolutionary time frame (e.g. that B1 is a descendent of A1), which can make it hard to see lineage turnover. Likewise, it can be difficult to identify co-circulating lineages with this system as sequences are assigned to later iterations. For example, although A1 may no longer appear to be circulating, enough variation may have accumulated to designate more recent sequences to descendant lineage A1.1. However, these limitations are primarily to do with the naming of lineages, and analyses should

be designed to incorporate the iterative nature of lineage designations, to represent the full extent of viral diversity. Additionally, the lineage designation steps implemented into the MAD DOG tool are not specific to RABV. The tool can therefore be used for any virus with just an input set of sequences and metadata. This may be valuable for other viruses with existing naming systems that need updating to show the appropriate definition for surveillance purposes, as the tool can incorporate existing naming systems and build from these.

One of the most recent and widely cited global phylogenomic analyses of RABV was published 5 years ago (Troupin et al. (2016)), using 321 WGS to capture and represent a level of geographic and temporal RABV diversity that was previously unexplored, covering all 6 major clades (Cosmopolitan, Asian, Africa-2, Africa-3, Arctic and Indian Subcontinent) [14]. However, since then publicly available RABV genomes have tripled and the way in which genomic surveillance is utilised has advanced from (usually) retrospective studies defining broad phylogenetic patterns to the scale of informing local control efforts in real-time. Therefore, we used this study as the baseline for designating coarse phylogenetic groups (using the existing clade names) but showcase the added value of our high resolution lineage designation tool for detailed epidemiological investigations. Sequence data has become increasingly accessible and we use 650 WGS just from the Cosmopolitan clade (over 50% of total dog related WGS available), to provide an improved characterisation of RABV diversity that supports our aim of interpreting sequence data for local and national surveillance. This approach follows from other local studies, for example, Brunker et al. (2015) identify 5 lineages circulating in Tanzania based on phylogenetic methods and genetic distances, showing their co-circulation and identifying historical introductions and human-mediated movement of infected animals [20]. However, different studies define and name lineages in different ways. A study from Cameroon used geographic area as part of its definition of viral lineages [34] while Talbi et al. [37] identify 8 “subtypes” within the Africa 2 clade (named A-H) defined by phylogenetic placement with Bayesian posterior support; thus splitting the subclade further, but not to a high resolution [37]. These examples highlight how the lack of a universal lineage definition and naming

system makes comparison between studies challenging, as the scale and method for lineage assignment and designation are not comparable.

An area that highlights the use of MAD DOG is the whole genome analysis of the Africa-1b subclade (AF1b), which is seen to be present in 9 African countries, with sequences spanning 1981-2018. Eleven of the 20 AF1b lineages have only been seen in Tanzania, 2 have only been seen in South Africa and the remainder have been detected in multiple countries (Table S1). This subclade has good definition due to the large number of sequences, with over 10% of these WGS from a large South African study [32] and several studies in Tanzania [11,18,36] that generated over 80% of the available AF1b WGS. When incorporating N gene data for AF1b, the limited diversity compared to WGS means additional N gene sequences do not translate into designated lineages (Table 1), as also evident in the ME2 and NEE subclades. The definition of designated lineages therefore depends both on the breadth of the data (the length of the sequences) and the depth of the data (number of sequences). Overall, a 25% decrease in definition, and in some instances up to 50%, is seen when using N gene sequences alone compared to WGS of the same sequence set. Additionally, higher bootstrap values are required for robust designations at this level due to the limited diversity. Despite the lower resolution provided by N gene data, the large volume of partial genome sequences makes their inclusion essential for representing previously identified diversity. As with the WGS designated lineages, those subclades not seen are due to having fewer than 10 sequences available (table 1).

A significant area of difference between the whole genome and N gene designations, that highlights the importance of using all available data, is found in the Africa-1a (AF1a) subclade. When only using whole genomes (21 WGS), the AF1a subclade splits into 2 lineages seen across multiple countries. With all the available N gene data (310 sequences) AF1a is split into 14 lineages, with distinct lineages present in Algeria and Morocco [37]. The first N gene sequence in the AF1a subclade is from 1982, whereas the first WGS is from 2 years later. Although the difference here is only two years, in other subclades, such as AM2b, the earliest available N gene sequences are from two decades earlier, thus excluding N gene sequences would exclude considerable historical data, and limit inference about historical patterns of dispersal.

Applying the MAD DOG lineage designation on a local scale allows in-country transmission routes and persistence of lineages to be seen at new depth. Tanzania is used as an example, which has a large number of WGS with detailed metadata. This provides greater inference to identify the uses and potential issues of this classification system on a local scale, such as local lineage dynamics and identification of areas needing improved vaccination efforts. 152 of the 205 Tanzanian sequences are from the Serengeti District, with 11 of the 16 Tanzanian lineages being detected here. This reflects greater sampling density enabling detection of circulating lineages, which also may be present elsewhere but remain undetected due to limited sampling. The sequence data point to the impact of improved dog vaccination in Serengeti district [40], with the apparent reduction in lineages from 10 in 2010 to just 3 by 2017. However, more sequences from 2017 onwards will be needed to determine whether lineage extinctions have truly occurred, or if limited ongoing sequencing has affected the detection of older lineages that are persisting in the district or elsewhere in the country.

While we provide proof of concept by applying MAD DOG to only the Cosmopolitan clade, the lineage classification and nomenclature system that we have developed can be easily extended to update all the RABV clades, and has potential to be extended to other viruses. The development of RABV-GLUE, and the incorporation of the updated classifications, provides an accessible platform for interpreting RABV sequence data with detailed genotyping that incorporates these high resolution lineage designations even for non-expert users. This functionality is important in rapid sequencing, and to inform policy, as it should allow detailed, accessible interpretations about a sequence to be given quickly, which may for example be used to identify sustained transmission or the source of an incursion. As well as the improved access to and interpretation of sequence data through RABV-GLUE, the increased resolution provided by this updated system can be used to better understand the spread and persistence of RABV and resulting lineage dynamics, as highlighted by the examples presented here. These tools should prove valuable for monitoring the progress of rabies control programmes as they strive to achieve the 2030 goal.

References

1. Pieracci EG. Vital Signs: Trends in Human Rabies Deaths and Exposures — United States, 1938–2018. MMWR Morb Mortal Wkly Rep. 2019;68. doi:10.15585/mmwr.mm6823e1
2. Hampson K, Coudeville L, Lembo T, Sambo M, Kieffer A, Attlan M, et al. Estimating the Global Burden of Endemic Canine Rabies. PLoS Negl Trop Dis. 2015;9: e0003709. doi:10.1371/journal.pntd.0003709
3. Taylor LH, Nel LH. Global epidemiology of canine rabies: past, present, and future prospects. Vet Med Res Rep. 2015;6: 361–371. doi:10.2147/VMRR.S51147
4. De Paula NS, Saraiva EA, Araújo IM, Nascimento KKG, Xavier DA, Santos KS, et al. Characterization of rabies post-exposure prophylaxis in a region of the eastern Amazon, state of Pará, Brazil, between 2000 and 2014. Zoonoses Public Health. 2018;65: 395–403. doi:10.1111/zph.12444
5. WHO | Rabies. [cited 12 Apr 2021]. Available: https://www.who.int/news-room/fact-sheets/detail/rabies
6. Bourhy H, Reynes J-M, Dunham EJ, Dacheux L, Larrous F, Huong VTQ, et al. The origin and phylogeography of dog rabies virus. J Gen Virol. 2008;89: 2673–2681. doi:10.1099/vir.0.2008/003913-0
7. Zhu S, Guo C. Rabies Control and Treatment: From Prophylaxis to Strategies with Curative Potential. Viruses. 2016;8. doi:10.3390/v8110279
8. WHO | United Against Rabies launches global plan to achieve zero rabies human deaths. [cited 12 Apr 2021]. Available: https://www.who.int/rabies/news/RUA-Rabies-launch-plan-achieve-zero-rabies-human-deaths-2030/en/
9. Lembo T, Hampson K, Haydon DT, Craft M, Dobson A, Dushoff J, et al. Exploring reservoir dynamics: a case study of rabies in the Serengeti ecosystem. J Appl Ecol. 2008;45: 1246–1257. doi:10.1111/j.1365-2664.2008.01468.x
10. Worsley-Tonks KEL, Escobar LE, Biek R, Castaneda-Guzman M, Craft ME, Streicker DG, et al. Using host traits to predict reservoir host species of rabies virus. PLoS Negl Trop Dis. 2020;14: e0008940. doi:10.1371/journal.pntd.0008940

11. Brunker K, Lemey P, Marston DA, Fooks AR, Lugelo A, Ngeleja C, et al. Landscape attributes governing local transmission of an endemic zoonosis: Rabies virus in domestic dogs. Mol Ecol. 2018;27: 773–788. doi:10.1111/mec.14470

12. Bourhy H, Nakouné E, Hall M, Nouvellet P, Lepelletier A, Talbi C, et al. Revealing the Micro-scale Signature of Endemic Zoonotic Disease Transmission in an African Urban Setting. PLOS Pathog. 2016;12: e1005525. doi:10.1371/journal.ppat.1005525

13. Nadin-Davis SA, Fehlner-Gardiner C. Origins of the arctic fox variant rabies viruses responsible for recent cases of the disease in southern Ontario. PLoS Negl Trop Dis. 2019;13: e0007699. doi:10.1371/journal.pntd.0007699

14. Troupin C, Dacheux L, Tanguy M, Sabeta C, Blanc H, Bouchier C, et al. Large-Scale Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in Multiple Carnivore Hosts. PLOS Pathog. 2016;12: e1006041. doi:10.1371/journal.ppat.1006041

15. Fischer S, Freuling CM, Müller T, Pfaff F, Bodenhofer U, Höper D, et al. Defining objective clusters for rabies virus sequences using affinity propagation clustering. PLoS Negl Trop Dis. 2018;12: e0006182. doi:10.1371/journal.pntd.0006182

16. CVR Bioinformatics. No Title. RABV-GLUE. Available: http://rabv-glue.cvr.gla.ac.uk/#/home

17. McElhinney LM, Marston DA, Freuling CM, Cragg W, Stankov S, Lalosević D, et al. Molecular diversity and evolutionary history of rabies virus strains circulating in the Balkans. J Gen Virol. 92: 2171–2180. doi:10.1099/vir.0.032748-0

18. Nadin-Davis SA, Colville A, Trewby H, Biek R, Real L. Application of high-throughput sequencing to whole rabies viral genome characterisation and its use for phylogenetic re-evaluation of a raccoon strain incursion into the province of Ontario. Virus Res. 2017;232: 123–133. doi:10.1016/j.virusres.2017.02.007

19. Gigante CM, Yale G, Condori RE, Costa NC, Long NV, Minh PQ, et al. Portable Rabies Virus Sequencing in Canine Rabies Endemic Countries Using the Oxford Nanopore MinION. Viruses. 2020;12: 1255. doi:10.3390/v12111255

20. Brunker K, Marston DA, Horton DL, Cleaveland S, Fooks AR, Kazwala R, et al. Elucidating the phylodynamics of endemic rabies virus in eastern Africa using whole-genome sequencing.

Virus Evol. 2015;1. doi:10.1093/ve/vev011

21. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5: 1403–1407. doi:10.1038/s41564-020-0770-5
22. Singer JB, Thomson EC, McLauchlan J, Hughes J, Gifford RJ. GLUE: a flexible software system for virus sequence data. BMC Bioinformatics. 2018;19: 532. doi:10.1186/s12859-018-2459-9
23. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014;2014: 2:2.
24. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066. doi:10.1093/nar/gkf436
25. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14: 587–589. doi:10.1038/nmeth.4285
26. Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30: 1188–1195. doi:10.1093/molbev/mst024
27. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4. doi:10.1093/ve/vex042
28. Kuzmin IV, Shi M, Orciari LA, Yager PA, Velasco-Villa A, Kuzmina NA, et al. Molecular Inferences Suggest Multiple Host Shifts of Rabies Viruses from Bats to Mesocarnivores in Arizona during 2001–2009. PLOS Pathog. 2012;8: e1002786. doi:10.1371/journal.ppat.1002786
29. Boland TA, McGuone D, Jindal J, Rocha M, Cumming M, Rupprecht CE, et al. Phylogenetic and Epidemiologic Evidence of Multiyear Incubation in Human Rabies. Ann Neurol. 2014;75: 155–160. doi:10.1002/ana.24016
30. Streicker DG, Winternitz JC, Satterfield DA, Condori-Condori RE, Broos A, Tello C, et al. Host-pathogen evolutionary signatures reveal dynamics and future invasions of vampire bat rabies. Proc Natl Acad Sci U S A. 2016;113: 10926–10931. doi:10.1073/pnas.1606587113
31. Nel LH, Rupprecht CE. Emergence of Lyssaviruses in the Old World: The Case of Africa. In:

Childs JE, Mackenzie JS, Richt JA, editors. Wildlife and Emerging Zoonotic Diseases: The Biology, Circumstances and Consequences of Cross-Species Transmission. Berlin, Heidelberg: Springer; 2007. pp. 161–193. doi:10.1007/978-3-540-70962-6_8

32. Sabeta C, Phahladira B, Marston DA, Wise EL, Ellis RJ, Fooks AR. Complete Genome Sequences of Six South African Rabies Viruses. Genome Announc. 2015;3: e01085-15. doi:10.1128/genomeA.01085-15

33. Sabeta C, Mohale D, Phahladira B, Ngoepe E, Van Schalkwyk A, Mogano K, et al. Complete Coding Sequences of 23 South African Domestic and Wildlife Rabies Viruses. Microbiol Resour Announc. 2020;9: e00621-20. doi:10.1128/MRA.00621-20

34. Maxmen A. One million coronavirus sequences: popular genome site hits mega milestone. Nature. 2021;593: 21–21. doi:10.1038/d41586-021-01069-w

35. GISAID. GISAID. Online; Available: https://www.gisaid.org/

36. Sadeuh-Mba SA, Momo JB, Besong L, Loul S, Njouom R. Molecular characterization and phylogenetic relatedness of dog-derived Rabies Viruses circulating in Cameroon between 2010 and 2016. PLoS Negl Trop Dis. 2017;11: e0006041. doi:10.1371/journal.pntd.0006041

37. Talbi C, Holmes EC, de Benedictis P, Faye O, Nakouné E, Gamatié D, et al. Evolutionary history and dynamics of dog rabies virus in western and central Africa. J Gen Virol. 90: 783–791. doi:10.1099/vir.0.007765-0

38. Brunker K, Jaswant G, Thumbi SM, Lushasi K, Lugelo A, Czupryna AM, et al. Rapid in-country sequencing of whole virus genomes to inform rabies elimination programmes. Wellcome Open Res. 2020;5: 3. doi:10.12688/wellcomeopenres.15518.2

39. Talbi C, Lemey P, Suchard MA, Abdelatif E, Elharrak M, Jalal N, et al. Phylodynamics and Human-Mediated Dispersal of a Zoonotic Virus. PLoS Pathog. 2010;6: e1001166. doi:10.1371/journal.ppat.1001166

40. Fitzpatrick MC, Hampson K, Cleaveland S, Meyers LA, Townsend JP, Galvani AP. Potential for Rabies Control through Dog Vaccination in Wildlife-Abundant Communities of Tanzania. PLoS Negl Trop Dis. 2012;6: e1796. doi:10.1371/journal.pntd.0001796

Supplementary Material

S1 Table - WGS Lineage Information

S2 Figure - N Gene Subset Designation

S3 Table - N Gene Lineage Information

S4 Table - Tanzania Lineage Designation

S5 Table - Tanzania Subset of WGS Lineage Information

S6 Figure - Global Sequence Length Maps and Time Series

S7 Figure - Tanzanian Sequence Length Time Series

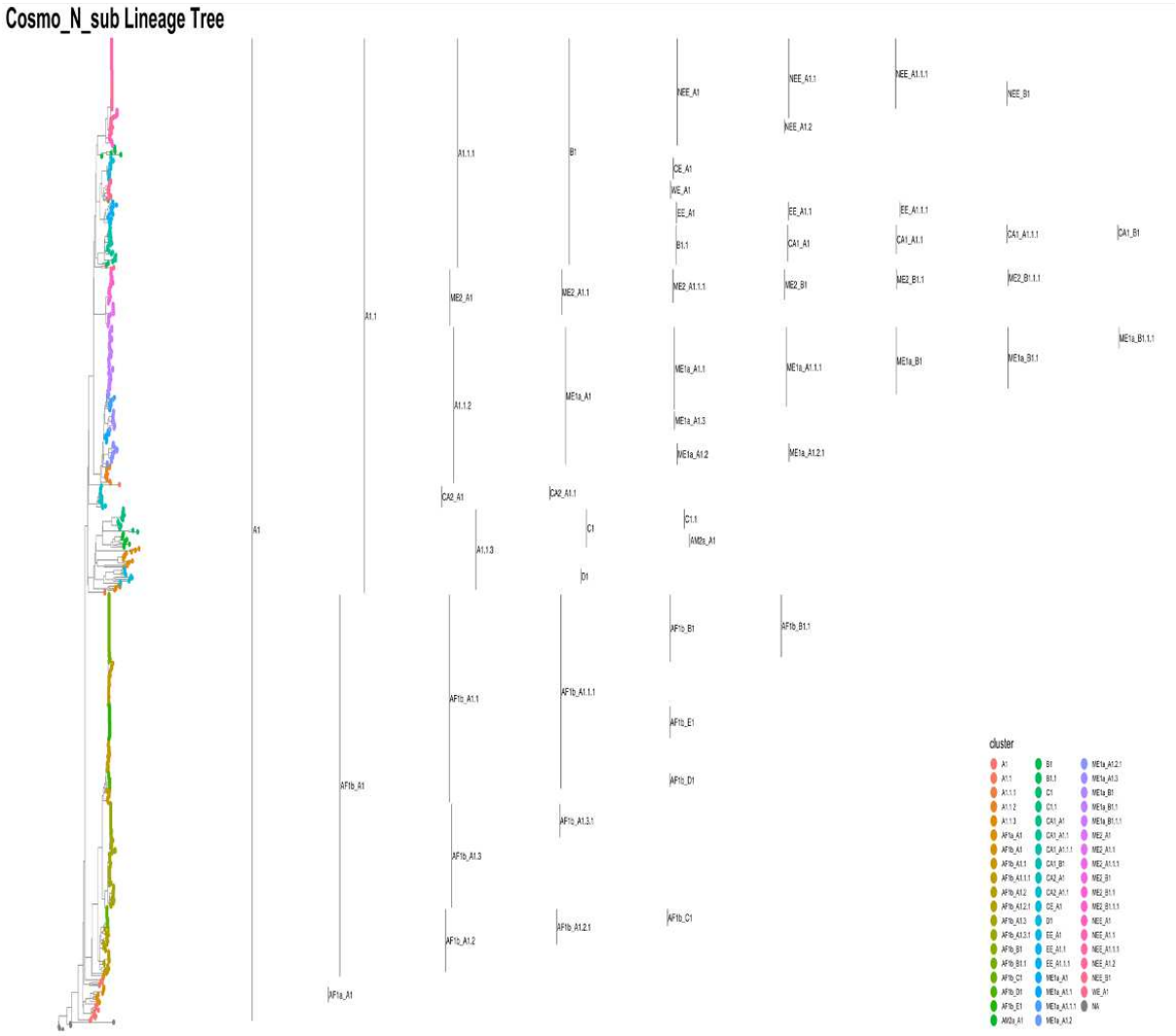
cluster	country	year_first	year_last	max_patristic_dist	mean_patristic_dist	n_seqs
A1	China	1931	2006	0.18	0.08	3
A1.1	Russia	2008	2014	0.18	0.08	2
A1.1.1	Gabon, Grenada, Nigeria	1986	2018	0.18	0.08	4
B1	Montenegro, Serbia	1978	1978	0.18	0.07	2
B1.1	Germany, Serbia	1972	1997	0.16	0.07	3
B1.1.1	Hungary, Russia	1991	2009	0.16	0.07	7
B1.2	Israel	1993	1993	0.17	0.06	2
C1	Madagascar	1986	2004	0.16	0.07	3
Cosmopolitan AF1a_A1	Algeria, Morocco, Tunisia	1989	2018	0.16	0.07	10
Cosmopolitan AF1a_A1.1	Ethiopia, Kenya, Somalia	1988	2018	0.16	0.07	11
Cosmopolitan AF1b_A1	Central African Republic, Tanzania, Zimbabwe	1991	2011	0.16	0.07	4
Cosmopolitan AF1b_A1.1	Rwanda, Tanzania	1994	2010	0.16	0.07	7
Cosmopolitan AF1b_A1.1.1	Tanzania, Uganda	2009	2018	0.16	0.07	4
Cosmopolitan AF1b_A1.2	South Africa, Tanzania, Zimbabwe	1981	2017	0.16	0.06	7
Cosmopolitan AF1b_A1.2.1	Mozambique, South Africa, Zimbabwe	1986	2012	0.16	0.07	5
Cosmopolitan AF1b_A1.3	Kenya, Tanzania	1996	2018	0.16	0.06	44
Cosmopolitan AF1b_A1.3.1	Tanzania	2010	2013	0.18	0.11	24
Cosmopolitan AF1b_A1.4	Namibia, South Africa	1992	2017	0.16	0.07	17
Cosmopolitan AF1b_B1	Tanzania	2010	2018	0.18	0.10	2
Cosmopolitan AF1b_B1.1	Tanzania	2008	2012	0.16	0.06	8
Cosmopolitan AF1b_B1.1.1	Tanzania	2004	2012	0.16	0.06	3
Cosmopolitan AF1b_C1	Tanzania	2003	2011	0.18	0.11	6

Cosmopolitan AF1b_C1.1	Tanzania	2009	2017	0.18	0.10	10
Cosmopolitan AF1b_D1	South Africa	2012	2015	0.16	0.06	2
Cosmopolitan AF1b_D1.1	South Africa	2015	2017	0.18	0.12	10
Cosmopolitan AF1b_E1	Tanzania	2010	2018	0.16	0.08	10
Cosmopolitan AF1b_E1.1	Tanzania	2010	2013	0.16	0.06	42
Cosmopolitan AF1b_F1	Tanzania	2015	2017	0.16	0.06	10
Cosmopolitan AF1b_F1.1	Tanzania	2011	2015	0.17	0.12	15
Cosmopolitan AF1b_G1	Tanzania	2010	2013	0.16	0.07	21
Cosmopolitan AM2a_A1	Mexico	1991	2018	0.16	0.07	10
Cosmopolitan CA1_A1	China, Poland, Russia	2008	2018	0.16	0.06	12
Cosmopolitan CA1_A1.1	Russia	2003	2003	0.16	0.06	2
Cosmopolitan CA1_A1.1.1	Iran	1974	2014	0.16	0.06	11
Cosmopolitan CA2_A1	Iran	1993	2015	0.16	0.07	2
Cosmopolitan CA2_A1.1	Iraq, Russia, Turkey	2001	2010	0.16	0.06	3
Cosmopolitan CA2_A1.1.1	Iran	2008	2014	0.16	0.06	10
Cosmopolitan CE_A1	Czechia, Germany, Poland	1991	2018	0.16	0.07	15
Cosmopolitan EE_A1	Bosnia and Herzegovina, Hungary, Slovenia	1986	1996	0.16	0.08	4
Cosmopolitan EE_A1.1	Hungary, Poland, Serbia	1986	2010	0.17	0.11	11
Cosmopolitan ME1a_A1	Turkey	2001	2001	0.16	0.06	2
Cosmopolitan ME1a_A1.1	Iran	1976	2013	0.16	0.06	8
Cosmopolitan ME1a_A1.1.1	Iran	2008	2018	0.16	0.06	3

Cosmopolitan ME1a_A1.1.2	Oman, Qatar, Saudi Arabia, United Arab Emirates	1987	2018	0.17	0.12	13
Cosmopolitan ME1a_A1.1.3	Iran	1984	2014	0.16	0.07	13
Cosmopolitan ME1a_A1.1.4	Israel, Turkey	1993	2001	0.16	0.07	10
Cosmopolitan ME1a_B1	Iran	2008	2014	0.16	0.07	3
Cosmopolitan ME1a_B1.1	Iran	2009	2014	0.16	0.08	4
Cosmopolitan ME1a_B1.1.1	Iran	1996	2014	0.17	0.11	4
Cosmopolitan ME1a_C1	Iran	2009	2014	0.18	0.12	8
Cosmopolitan ME1a_C1.1	Iran	2008	2014	0.16	0.08	7
Cosmopolitan ME1a_C1.1.1	Iran	2008	2014	0.16	0.06	11
Cosmopolitan ME1a_C1.2	Iran	2009	2015	0.16	0.07	15
Cosmopolitan ME2_A1	Turkey	1993	2001	0.17	0.07	7
Cosmopolitan ME2_A1.1	Iran, Turkey	1989	2014	0.16	0.06	8
Cosmopolitan ME2_A1.1.1	Turkey	2001	2001	0.16	0.06	2
Cosmopolitan ME2_B1	Turkey	1999	2015	0.16	0.06	9
Cosmopolitan ME2_B1.1	Turkey	2009	2010	0.16	0.06	2
Cosmopolitan ME2_B1.1.1	Turkey	2006	2014	0.16	0.06	10
Cosmopolitan NEE_A1	Estonia	1991	1993	0.16	0.06	4
Cosmopolitan NEE_A1.1	Poland	1996	1997	0.16	0.06	4
Cosmopolitan NEE_A1.1.1	Hungary, Poland	2010	2017	0.16	0.07	6
Cosmopolitan NEE_A1.1.2	Estonia, Finland, Poland, Russia	1986	2018	0.16	0.07	10

Cosmopolitan NEE_B1	Hungary	2013	2013	0.17	0.12	2
Cosmopolitan NEE_B1.1	Hungary	2013	2014	0.16	0.08	28
Cosmopolitan NEE_B1.2	Hungary	2013	2014	0.15	0.08	17
Cosmopolitan WE_A1	Belgium, Bosnia and Herzegovina, France, Germany, Serbia, Slovenia	1986	1998	0.15	0.07	13
Cosmopolitan_A1	China, Ethiopia, United States	1948	2018	0.16	0.07	14
D1	United States	1982	2012	0.16	0.07	4
D1.1	Egypt, Israel, Mexico, United States	1950	2015	0.16	0.07	13
D1.1.1	United States	2009	2009	0.16	0.07	2
D1.1.2	Brazil, United States	1986	2018	0.16	0.07	11
E1	Azerbaijan, Tajikistan	2002	2012	0.16	0.08	2

S1 Table - WGS Lineage Information. *Details of the 73 designated lineages from 650 whole genome (10,000nt) rabies virus sequences from the Cosmopolitan clade, obtained from RABV-GLUE. Includes all countries sequences assigned each lineage have been seen in, the first and most recent collection years of those sequences, the maximum and mean patristic distance of all sequences in each lineage and the lineages descended from it, and the number of sequences assigned to each lineage.*



S2 Figure - N Gene Subset Designation. *Lineage designation of the N genes (1300nt) of the 650 available whole genome rabies virus sequences from the Cosmopolitan clade, obtained from RABV-GLUE.*

cluster	country	year_first	year_last	max_patristic_dist	mean_patristic_dist	n_seqs
A1	Russia	2008	2014	0.19	0.09	3
A1.1	Cameroon, Equatorial Guinea, Gabon, Nigeria	1986	2018	0.19	0.09	22
A1.1.1	Grenada, Madagascar	1986	2012	0.19	0.09	6
A1.2	Somalia	1993	1993	0.19	0.08	2
A1.2.1	Ethiopia, Kenya	1987	2018	0.19	0.08	13
A1.3	Tunisia	1986	2016	0.17	0.08	6
A1.3.1	Tunisia	1986	2010	0.14	0.07	10
AF1a_A1	Morocco	1986	2010	0.16	0.08	8
AF1a_A1.1	Morocco, Spain	1985	2010	0.16	0.08	4
AF1a_A1.1.1	Spain	1996	2010	0.17	0.08	3
AF1a_A1.1.2	Spain	1993	2010	0.19	0.10	7
AF1a_A1.1.3	Morocco	1991	2010	0.19	0.11	2
AF1a_B1	NA	1995	2010	0.19	0.09	18
AF1a_B1.1	NA	2010	2010	0.19	0.09	4
AF1a_B1.1.1	NA	2010	2010	0.19	0.09	17
AF1a_B1.2	NA	2010	2010	0.14	0.07	2
AF1a_B1.2.1	NA	2010	2010	0.17	0.08	3
AF1a_B1.3	NA	2010	2010	0.15	0.08	3
AF1a_B1.3.1	NA	2010	2010	0.15	0.08	13
AF1a_B1.4	Morocco, Spain	1993	2010	0.17	0.08	19
AF1a_B1.5	NA	2010	2010	0.16	0.09	21
AF1a_B1.6	NA	2010	2010	0.17	0.08	11
AF1a_C1	Morocco	1990	2010	0.19	0.11	4
AF1a_C1.1	Morocco	2004	2010	0.16	0.08	10
AF1a_D1	Morocco	1991	2010	0.14	0.07	4
AF1a_D1.1	Morocco	2009	2010	0.16	0.10	13
AF1a_E1	NA	2010	2010	0.16	0.08	4

AF1a_E1.1	NA	2010	2010	0.16	0.08	15
AF1a_F1	Algeria	2010	2015	0.18	0.11	58
AF1a_F1.1	NA	2010	2010	0.16	0.08	13
AF1a_G1	NA	2010	2010	0.15	0.08	12
AF1a_H1	Algeria, Tunisia	1986	2010	0.19	0.11	15
AF1a_I1	NA	2010	2010	0.19	0.11	15
AF1a_NA	NA	1995	2018	0.14	0.07	12
AF1b_A1	Tanzania	1995	2011	0.16	0.08	3
AF1b_A1.1	Central African Republic, Uganda	1992	2011	0.16	0.08	4
AF1b_A1.1.1	Central African Republic, Democratic Republic of the Congo, Zambia	1989	2011	0.18	0.11	138
AF1b_B1	Namibia, South Africa, Zimbabwe	1990	2016	0.18	0.11	19
AF1b_B1.1	South Africa, Tanzania, Zimbabwe	1981	2017	0.16	0.08	16
AF1b_B1.1.1	Mozambique, South Africa	1986	2012	0.16	0.08	5
AF1b_B1.2	South Africa	1991	2005	0.14	0.07	5
AF1b_B1.2.1	Namibia, South Africa	1992	2017	0.16	0.08	12
AF1b_C1	Burundi, Rwanda, Tanzania	1990	2010	0.15	0.08	9
AF1b_C1.1	Uganda	2009	2012	0.16	0.08	3
AF1b_C1.1.1	Tanzania	1998	2018	0.16	0.08	61
AF1b_D1	Kenya, Tanzania	1994	2018	0.17	0.11	79
AF1b_D1.1	Tanzania	1997	2013	0.17	0.11	10
AF1b_E1	Tanzania	2004	2010	0.16	0.08	2
AF1b_E1.1	Tanzania	2009	2017	0.16	0.08	10
AF1b_F1	Tanzania	2010	2013	0.14	0.07	42
AF1b_G1	Tanzania	2010	2013	0.14	0.07	21
AF1b_H1	South Africa	2012	2017	0.14	0.08	12
AF4_A1	Egypt, Israel	1950	2009	0.16	0.08	13

AM1_A1	United States	1984	2012	0.15	0.08	4
AM1_A1.1	United States	1982	2012	0.16	0.08	3
AM1_A1.1.1	Canada, United States	2000	2009	0.16	0.08	17
AM2a_A1	Colombia, Cuba, Mexico	1995	2014	0.17	0.11	16
AM2a_A1.1	Cuba	2000	2002	0.18	0.11	4
AM2a_A1.1.1	Cuba	2000	2002	0.19	0.12	6
AM2a_A1.2	Cuba	2000	2002	0.16	0.10	11
AM2a_B1	Cuba	2000	2002	0.14	0.07	13
AM3a_A1	Brazil	1986	2007	0.13	0.07	4
AM3a_A1.1	Brazil	2006	2007	0.13	0.07	9
AM3a_A1.1.1	Brazil	2007	2007	0.14	0.07	4
AM3a_A1.1.2	Brazil	2007	2007	0.14	0.08	3
AM3a_A1.2	Brazil	1995	2007	0.16	0.08	3
AM3a_A1.2.1	Brazil	2002	2007	0.16	0.08	5
AM3a_A1.3	Brazil, Peru, United States	1999	2011	0.16	0.08	10
AM3a_A1.4	Brazil	2006	2009	0.16	0.08	22
AM3a_B1	Brazil	2002	2007	0.17	0.11	7
AM3a_B1.1	Brazil	2002	2007	0.18	0.12	20
AM3a_B1.1.1	Brazil	2002	2007	0.17	0.11	10
AM3a_C1	Brazil	2006	2012	0.15	0.08	13
AM3a_D1	Brazil	2006	2012	0.15	0.08	10
AM3b_A1	Brazil	2006	2007	0.15	0.08	7
AM3b_A1.1	Brazil	2006	2006	0.19	0.11	6
AM3b_A1.1.1	Brazil	2005	2012	0.16	0.10	11
AM4_A1	United States	1974	1997	0.15	0.08	5
AM4_A1.1	United States	1994	2016	0.13	0.07	12
B1	Montenegro, Peru, Serbia, United States	1978	2016	0.13	0.07	7
B1.1	Mexico	2000	2015	0.13	0.07	8

B1.1.1	Brazil	1986	2018	0.14	0.07	5
B1.1.2	Mexico, United States	1986	2016	0.14	0.07	7
B1.2	Puerto Rico	1997	2016	0.15	0.08	5
B1.2.1	Puerto Rico	1996	2006	0.17	0.08	10
B1.3	Brazil, Peru	2004	2016	0.16	0.08	11
C1	Germany, Hungary	1991	2007	0.16	0.08	7
C1.1	Azerbaijan, Georgia, Tajikistan	2002	2015	0.16	0.08	7
C1.1.1	Russia	2004	2013	0.15	0.08	5
C1.1.2	NA	2004	2004	0.15	0.08	2
C1.2	France	1994	1999	0.15	0.08	2
C1.2.1	Bosnia and Herzegovina, Italy, Serbia, Slovenia	1986	2009	0.15	0.08	10
CA1_A1	Russia	2001	2008	0.16	0.08	7
CA1_A1.1	Mongolia, Russia	2002	2018	0.16	0.08	8
CA1_A1.1.1	Mongolia	2005	2008	0.17	0.11	12
CA1_A1.2	NA	2004	2004	0.18	0.11	3
CA1_A1.2.1	Kazakhstan, Mongolia, Russia	2002	2014	0.18	0.11	11
CA1_B1	China, Russia)	2011	2018	0.17	0.11	4
CA1_B1.1	Russia	2017	2017	0.18	0.10	3
CA1_B1.1.1	Russia	2017	2018	0.15	0.08	10
CA1_B1.2	China, Russia	2013	2019	0.15	0.09	22
CA1_B1.3	Mongolia, Russia	2017	2018	0.15	0.08	10
CA1_C1	China, Mongolia	2006	2018	0.15	0.08	17
CA1_C1.1	China, Mongolia	2005	2018	0.15	0.08	13
CA1_D1	Russia	2011	2012	0.15	0.08	11
CA2_A1	Iran, Iraq, Turkey	1993	2020	0.14	0.08	9
CA2_A1.1	Georgia, Russia	2004	2016	0.15	0.09	8
CA2_A1.1.1	Georgia, Turkey	2001	2016	0.16	0.09	9
CA2_A1.2	Iran	2008	2014	0.19	0.11	10

CA2_B1	Georgia	2015	2015	0.18	0.11	6
CA2_B1.1	Russia	2008	2008	0.16	0.10	2
CA2_B1.1.1	Georgia	2015	2015	0.16	0.10	5
CA2_B1.2	Georgia	2015	2016	0.19	0.12	27
CA2_C1	Georgia	2015	2016	0.18	0.10	12
CA3_A1	Hungary	1991	2009	0.16	0.08	6
CA3_A1.1	Romania	2004	2011	0.13	0.07	4
CA3_A1.1.1	Russia	2009	2014	0.15	0.08	4
CA3_B1	Russia	2009	2016	0.14	0.07	10
CA3_B1.1	Russia	2008	2016	0.14	0.07	7
CA3_B1.1.1	Russia	2012	2014	0.14	0.07	10
CE_A1	Czechia, Germany, Poland	1991	2018	0.13	0.07	18
D1	Israel, Jordan, Turkey	1993	2001	0.13	0.07	8
D1.1	Iran	2008	2013	0.13	0.07	2
D1.1.1	Iran	2008	2018	0.13	0.07	8
D1.2	Saudi Arabia, Turkey	1987	2001	0.13	0.07	3
D1.2.1	Israel, Jordan, Qatar, Saudi Arabia	1997	2018	0.14	0.07	13
D1.2.2	Oman, Saudi Arabia, United Arab Emirates	1987	2018	0.14	0.07	11
D1.3	Iran	1984	2014	0.14	0.07	15
D1.4	Israel	1993	2005	0.14	0.07	13
D1.5	Iran	1976	2013	0.14	0.07	11
E1	United States	1950	2016	0.14	0.08	6
E1.1	China, Ethiopia, France, India	1976	2017	0.15	0.09	12
E1.1.1	China, France, India, United States	1988	2018	0.15	0.08	11
E1.2	China, France, United States	1939	2013	0.15	0.08	11
EE_A1	Bosnia and Herzegovina, Hungary, Montenegro, Serbia, Slovenia	1986	2000	0.17	0.08	14

EE_A1.1	Poland	1992	1994	0.16	0.08	2
EE_A1.1.1	Hungary, Serbia	1977	1993	0.16	0.08	5
EE_B1	Serbia	1977	1986	0.16	0.08	6
EE_B1.1	Hungary, Serbia	1986	2010	0.15	0.08	10
EE_B1.1.1	Serbia	1998	2000	0.15	0.08	10
F1	Mexico	1991	2018	0.15	0.08	6
F1.1	Mexico	2000	2011	0.15	0.08	28
F1.2	Mexico	1990	2010	0.17	0.08	18
G1	Brazil	2005	2010	0.15	0.08	9
G1.1	Brazil	2006	2011	0.16	0.08	10
G1.2	Brazil	2005	2010	0.16	0.08	18
H1	Brazil	2007	2009	0.15	0.08	3
H1.1	Brazil	2005	2007	0.15	0.08	14
I1	Iran, Turkey	1989	2014	0.16	0.08	8
I1.1	Israel	2004	2011	0.16	0.08	22
ME1a_A1	Israel	1997	1998	0.16	0.08	3
ME1a_A1.1	Israel	1993	1998	0.16	0.09	11
ME1a_A1.1.1	Israel	1996	2004	0.17	0.10	13
ME1a_A1.1.2	Israel	1996	2004	0.17	0.10	10
ME1a_A1.2	Israel, Jordan, Lebanon	1996	2004	0.17	0.11	14
ME1a_B1	Israel	1995	2001	0.18	0.11	10
ME2_A1	Turkey	2001	2001	0.17	0.11	2
ME2_A1.1	Turkey	1999	2015	0.17	0.11	7
ME2_A1.1.1	Turkey	1999	2004	0.17	0.11	2
ME2_B1	Turkey	2006	2014	0.19	0.12	12
NA.1	Iran	2005	2014	0.18	0.12	27
NA.1.1	Iran	2009	2015	0.18	0.12	15
NA.2	Grenada	2011	2013	0.17	0.11	6

NA.2.1	Grenada	2011	2013	0.17	0.11	10
NA.3	Brazil	2006	2010	0.17	0.11	12
NA.4	Brazil	2007	2007	0.17	0.11	10
NA.5	China, Kazakhstan, Russia	2004	2018	0.18	0.12	11
NEE_A1	Estonia, Poland	1991	1997	0.15	0.08	8
NEE_A1.1	Hungary, Poland	2010	2017	0.15	0.08	6
NEE_A1.1.1	Hungary	2013	2014	0.15	0.08	30
NEE_A1.2	Finland	1988	1993	0.15	0.08	2
NEE_A1.2.1	Estonia, Poland, Russia	1986	2018	0.15	0.09	11
NEE_B1	Hungary	2013	2014	0.15	0.08	17
WE_A1	Belgium, France	1994	1999	0.15	0.08	4
WE_A1.1	France	1991	1999	0.15	0.08	10

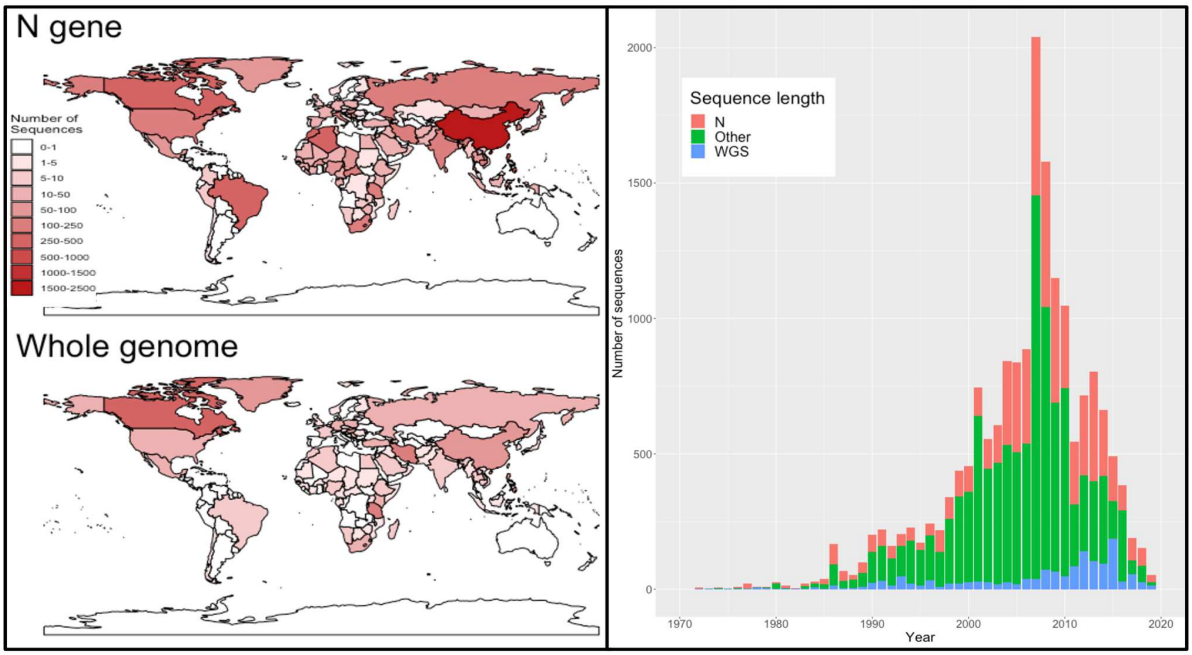
S3 Table - N Gene Lineage Information. *Details of the 168 designated lineages from 2126 N gene (1300nt) rabies virus sequences from the Cosmopolitan clade, obtained from RABV-GLUE. Includes all countries sequences assigned each lineage have been seen in, the first and most recent collection years of those sequences, the maximum and mean patristic distance of all sequences in each lineage and the lineages descended from it, and the number of sequences assigned to each lineage. Where country = NA, no information is publicly available for sequence locations.*

cluster	*place*	*year_first*	*year_last*	*max_patristic_dist*	*mean_patristic_dist*	*n_seqs*
Cosmopolitan AF1b_A1	*Morogoro region, Serengeti District*	*2009*	*2017*	*0.07*	*0.02*	*7*
Cosmopolitan AF1b_A1.1	*Arusha Region, Mtwara region, Pemba island, Serengeti District*	*2003*	*2011*	*0.07*	*0.02*	*4*
Cosmopolitan AF1b_A1.1.1	*Pwani region*	*2010*	*2018*	*0.07*	*0.02*	*2*
Cosmopolitan AF1b_A1.1.2	*Arusha Region, Chake chake, Dar es Salaam, Iringa region, Kusini Unguja, Lindi region, Morogoro region, Pemba island, Serengeti District*	*1996*	*2017*	*0.07*	*0.02*	*37*
Cosmopolitan AF1b_B1	*Serengeti District*	*2004*	*2012*	*0.07*	*0.02*	*11*
Cosmopolitan AF1b_B1.1	*Morogoro region, Serengeti District*	*2010*	*2013*	*0.07*	*0.02*	*7*
Cosmopolitan AF1b_B1.1.1	*NA*	*2017*	*2018*	*0.07*	*0.02*	*3*
Cosmopolitan AF1b_B1.2	*Serengeti District*	*2010*	*2013*	*0.07*	*0.02*	*21*
Cosmopolitan AF1b_B1.3	*Serengeti District*	*2011*	*2017*	*0.07*	*0.02*	*28*
Cosmopolitan AF1b_C1	*Serengeti District*	*2012*	*2013*	*0.07*	*0.02*	*2*
Cosmopolitan AF1b_C1.1	*Serengeti District*	*2011*	*2012*	*0.07*	*0.02*	*5*
Cosmopolitan AF1b_C1.1.1	*Morogoro region, Serengeti District*	*2010*	*2013*	*0.07*	*0.02*	*35*
Cosmopolitan AF1b_D1	*Dar es Salaam, Morogoro region, Serengeti District*	*2010*	*2012*	*0.07*	*0.02*	*14*
Cosmopolitan AF1b_E1	*Serengeti District*	*2010*	*2013*	*0.07*	*0.02*	*23*

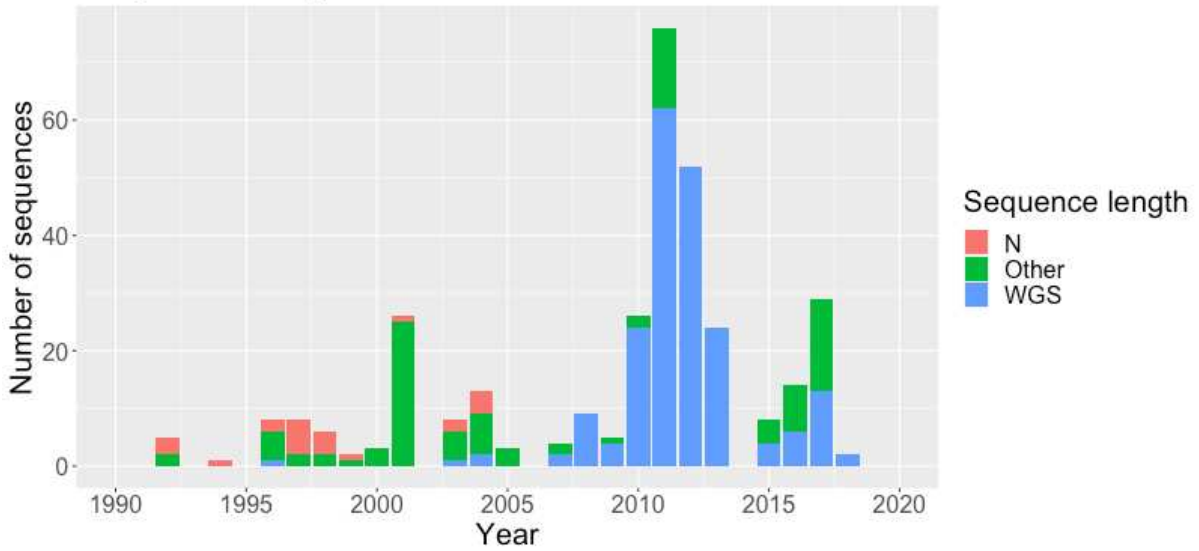
S4 Table - Tanzania Lineage Designation. *Details of the 14 designated lineages from whole genome (10000nt) rabies virus sequences from Tanzania, obtained from RABV-GLUE. Includes all places sequences assigned each lineage have been seen in, the first and most recent collection years of those sequences, the maximum and mean patristic distance of all sequences in each lineage and the lineages descended from it, and the number of sequences assigned to each lineage. Where place = NA, no information is publicly available for sequence locations.*

cluster	*country*	*year_first*	*year_last*	*n_seqs*
Cosmopolitan AF1b_A1	*Chake chake, Kusini Unguja*	*2010*	*2011*	*2*
Cosmopolitan AF1b_A1.1	*Morogoro region, Pemba island*	*2008*	*2010*	*6*
Cosmopolitan AF1b_A1.1.1	*Serengeti District*	*2011*	*2013*	*2*
Cosmopolitan AF1b_A1.2	*Dar es Salaam, Iringa region, Lindi region*	*2010*	*2011*	*4*
Cosmopolitan AF1b_A1.3	*Arusha Region, Dar es Salaam, Iringa region, Morogoro region, Serengeti District*	*1996*	*2017*	*37*
Cosmopolitan AF1b_A1.3.1	*Serengeti District*	*2010*	*2013*	*24*
Cosmopolitan AF1b_B1	*Pwani region*	*2010*	*2018*	*2*
Cosmopolitan AF1b_B1.1	*Serengeti District*	*2008*	*2012*	*8*
Cosmopolitan AF1b_B1.1.1	*Serengeti District*	*2004*	*2012*	*3*
Cosmopolitan AF1b_C1	*Arusha Region, Morogoro region, Mtwara region, Pemba island, Serengeti District*	*2003*	*2011*	*6*
Cosmopolitan AF1b_C1.1	*Lindi region, Pemba island, Serengeti District*	*2009*	*2017*	*10*
Cosmopolitan AF1b_E1	*Morogoro region, Serengeti District*	*2010*	*2018*	*10*
Cosmopolitan AF1b_E1.1	*Morogoro region, Serengeti District*	*2010*	*2013*	*42*
Cosmopolitan AF1b_F1	*NA*	*2015*	*2017*	*10*
Cosmopolitan AF1b_F1.1	*Serengeti District*	*2011*	*2015*	*15*
Cosmopolitan AF1b_G1	*Serengeti District*	*2010*	*2013*	*21*

S5 Table - Tanzania Subset of WGS Lineage Information. *Details of the 16 global MAD DOG whole genome lineages (10000nt) detected in Tanzania, obtained from RABV-GLUE. Includes all places sequences assigned each lineage have been seen in, the first and most recent collection years of those sequences and the number of sequences assigned to each lineage. Where place = NA, no information is publicly available for sequence locations.*



S6 Figure - Global Sequence Length Maps and Time Series. Left: *Map of collection locations of all available whole genome and N gene sequences.* ***Right:*** *Time series of number of whole genome, N gene, and shorter sequences collected.*



S7 Figure - Tanzanian Sequence Length Time Series. *Timeseries of the number of whole genome, N gene, and shorter partial genome sequences from Tanzania available on RABV-GLUE.*