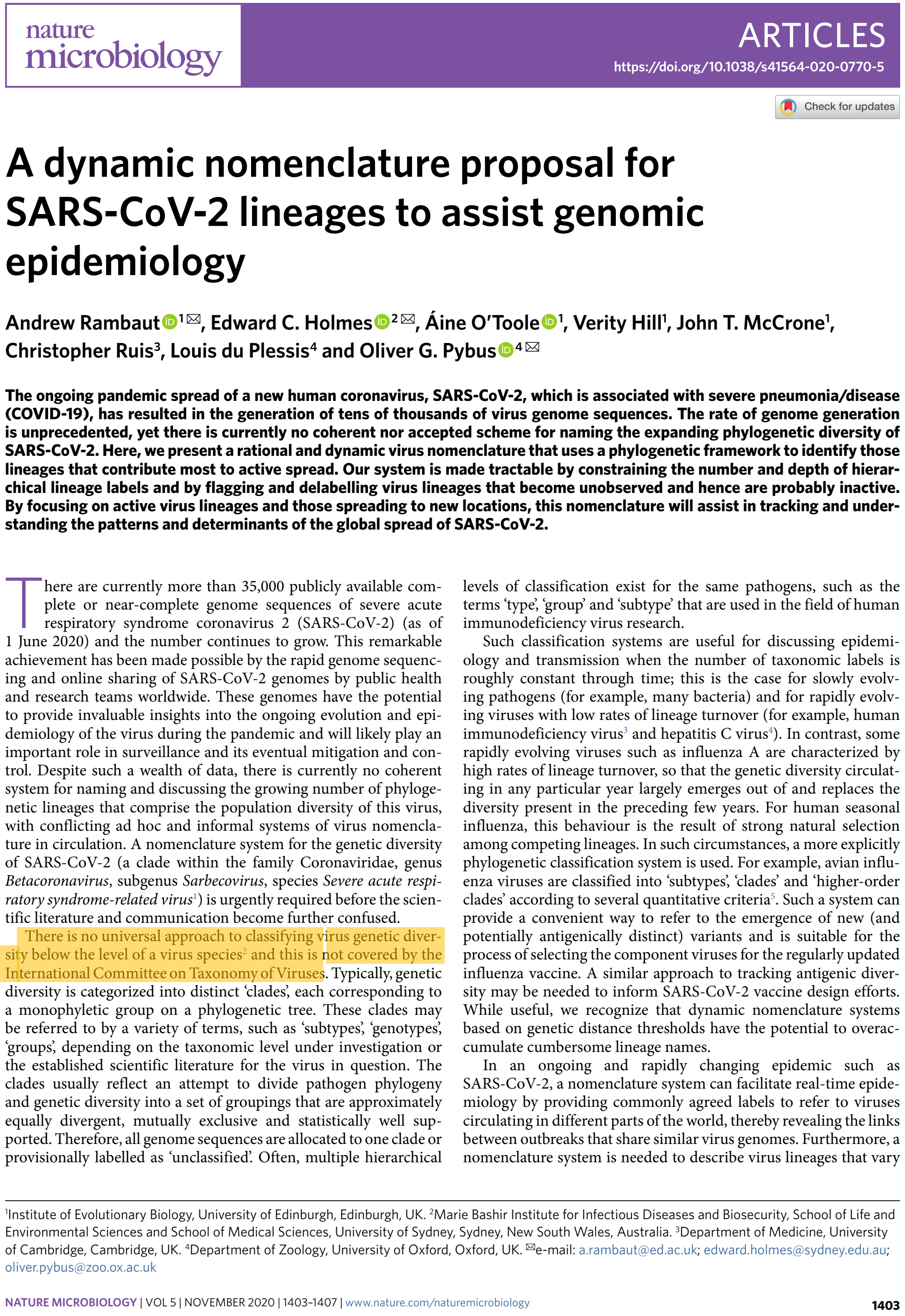
Pango beyond the SARS-CoV-2 pandemic

Rachel Colquhoun, Áine O’Toole, Oliver Pybus and Andrew Rambaut

Viral Subspecies Classification Workshop, BV-BRC

8-10th April 2024



nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-020-0770-5

Check for updates

A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology

Andrew Rambaut[1]✉, Edward C. Holmes[2]✉, Áine O'Toole[1], Verity Hill[1], John T. McCrone[1], Christopher Ruis[3], Louis du Plessis[4] and Oliver G. Pybus[4]✉

The ongoing pandemic spread of a new human coronavirus, SARS-CoV-2, which is associated with severe pneumonia/disease (COVID-19), has resulted in the generation of tens of thousands of virus genome sequences. The rate of genome generation is unprecedented, yet there is currently no coherent nor accepted scheme for naming the expanding phylogenetic diversity of SARS-CoV-2. Here, we present a rational and dynamic virus nomenclature that uses a phylogenetic framework to identify those lineages that contribute most to active spread. Our system is made tractable by constraining the number and depth of hierarchical lineage labels and by flagging and delabelling virus lineages that become unobserved and hence are probably inactive. By focusing on active virus lineages and those spreading to new locations, this nomenclature will assist in tracking and understanding the patterns and determinants of the global spread of SARS-CoV-2.

There are currently more than 35,000 publicly available complete or near-complete genome sequences of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (as of 1 June 2020) and the number continues to grow. This remarkable achievement has been made possible by the rapid genome sequencing and online sharing of SARS-CoV-2 genomes by public health and research teams worldwide. These genomes have the potential to provide invaluable insights into the ongoing evolution and epidemiology of the virus during the pandemic and will likely play an important role in surveillance and its eventual mitigation and control. Despite such a wealth of data, there is currently no coherent system for naming and discussing the growing number of phylogenetic lineages that comprise the population diversity of this virus, with conflicting ad hoc and informal systems of virus nomenclature in circulation. A nomenclature system for the genetic diversity of SARS-CoV-2 (a clade within the family Coronaviridae, genus *Betacoronavirus*, subgenus *Sarbecovirus*, species *Severe acute respiratory syndrome-related virus*[1]) is urgently required before the scientific literature and communication become further confused.

There is no universal approach to classifying virus genetic diversity below the level of a virus species[2] and this is not covered by the International Committee on Taxonomy of Viruses. Typically, genetic diversity is categorized into distinct 'clades', each corresponding to a monophyletic group on a phylogenetic tree. These clades may be referred to by a variety of terms, such as 'subtypes', 'genotypes', 'groups', depending on the taxonomic level under investigation or the established scientific literature for the virus in question. The clades usually reflect an attempt to divide pathogen phylogeny and genetic diversity into a set of groupings that are approximately equally divergent, mutually exclusive and statistically well supported. Therefore, all genome sequences are allocated to one clade or provisionally labelled as 'unclassified'. Often, multiple hierarchical levels of classification exist for the same pathogens, such as the terms 'type', 'group' and 'subtype' that are used in the field of human immunodeficiency virus research.

Such classification systems are useful for discussing epidemiology and transmission when the number of taxonomic labels is roughly constant through time; this is the case for slowly evolving pathogens (for example, many bacteria) and for rapidly evolving viruses with low rates of lineage turnover (for example, human immunodeficiency virus[3] and hepatitis C virus[4]). In contrast, some rapidly evolving viruses such as influenza A are characterized by high rates of lineage turnover, so that the genetic diversity circulating in any particular year largely emerges out of and replaces the diversity present in the preceding few years. For human seasonal influenza, this behaviour is the result of strong natural selection among competing lineages. In such circumstances, a more explicitly phylogenetic classification system is used. For example, avian influenza viruses are classified into 'subtypes', 'clades' and 'higher-order clades' according to several quantitative criteria[5]. Such a system can provide a convenient way to refer to the emergence of new (and potentially antigenically distinct) variants and is suitable for the process of selecting the component viruses for the regularly updated influenza vaccine. A similar approach to tracking antigenic diversity may be needed to inform SARS-CoV-2 vaccine design efforts. While useful, we recognize that dynamic nomenclature systems based on genetic distance thresholds have the potential to overaccumulate cumbersome lineage names.

In an ongoing and rapidly changing epidemic such as SARS-CoV-2, a nomenclature system can facilitate real-time epidemiology by providing commonly agreed labels to refer to viruses circulating in different parts of the world, thereby revealing the links between outbreaks that share similar virus genomes. Furthermore, a nomenclature system is needed to describe virus lineages that vary

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK. [2]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, University of Sydney, Sydney, New South Wales, Australia. [3]Department of Medicine, University of Cambridge, Cambridge, UK. [4]Department of Zoology, University of Oxford, Oxford, UK. ✉e-mail: a.rambaut@ed.ac.uk; edward.holmes@sydney.edu.au; oliver.pybus@zoo.ox.ac.uk

NATURE MICROBIOLOGY | VOL 5 | NOVEMBER 2020 | 1403–1407 | www.nature.com/naturemicrobiology

"There is no universal approach to classifying virus genetic diversity below the level of a virus species[2], and this is not covered by the International Committee on Taxonomy of Viruses (ICTV)"

Authors

Áine O’Toole

Oliver Pybus

Andrew Rambaut

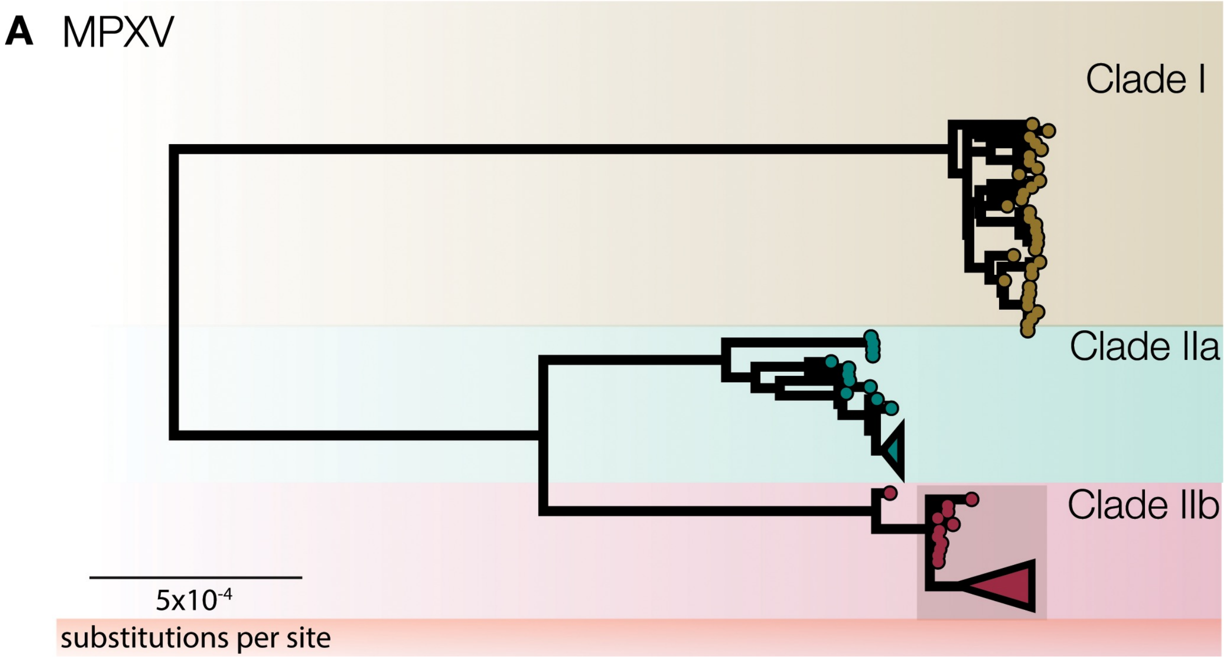
Overview

- Pango Lineages for MPXV
- Generalized definitions
- Challenges

pangō
verb (latin)
I fix, set or record

Case Study: MPXV

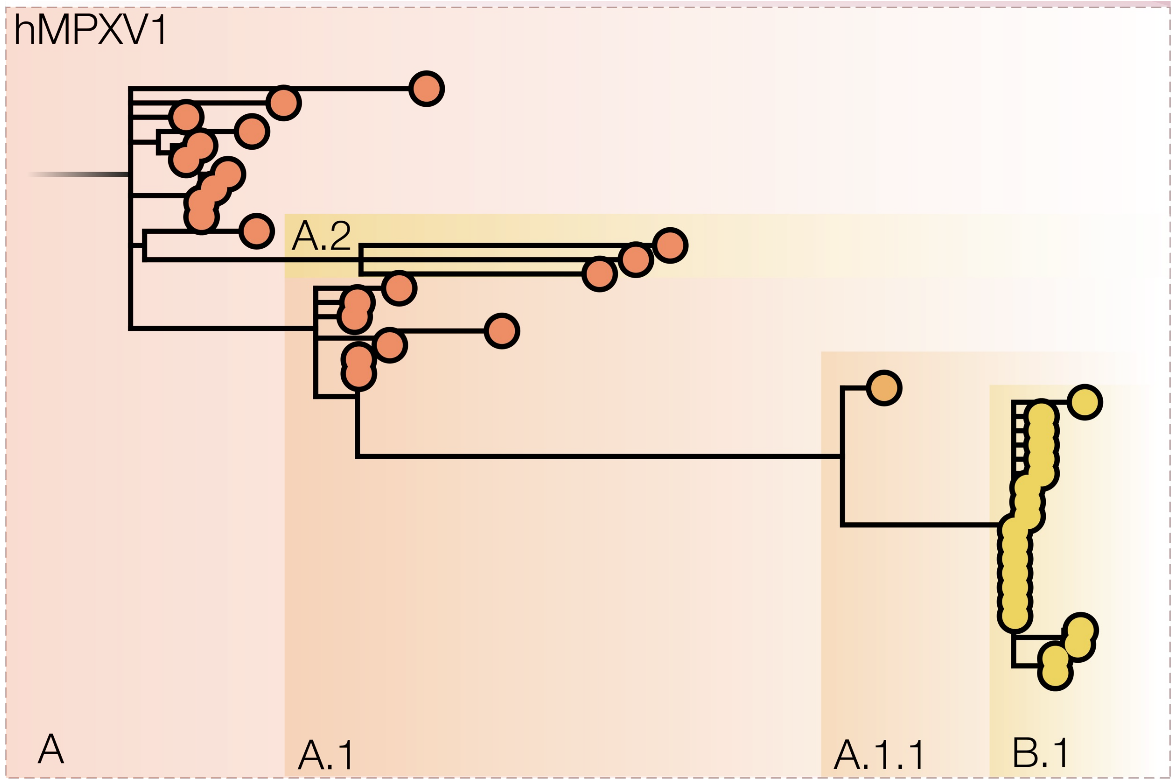
2022



MPXV

Happi C, Adetifa I, Mbala P, Njouom R, Nakoune E, Happi A, et al. (2022) Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus. PLoS Biol 20(8): e3001769. https://doi.org/10.1371/journal.pbio.3001769

2022



hMPXV1

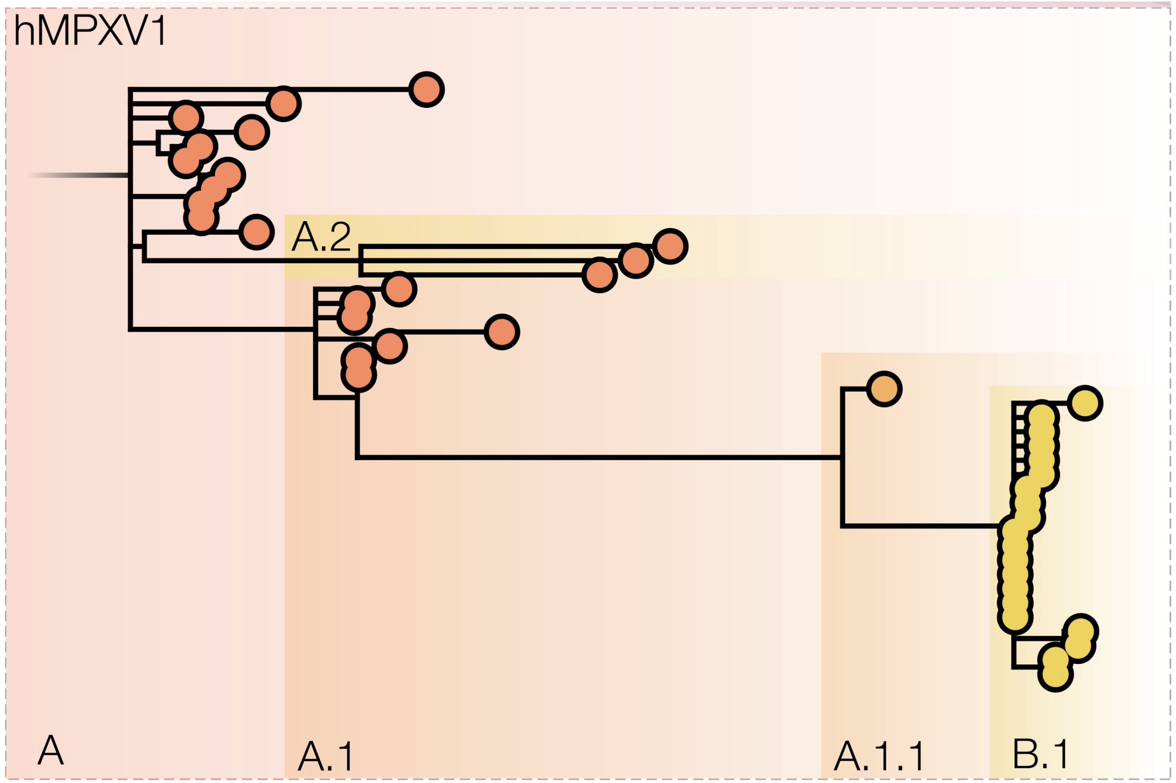
Fine-scale

Alias introduced earlier

A.2 and B.1 both associated with samples from 2022

Happi et al, PLoS Biol (2022)

2022

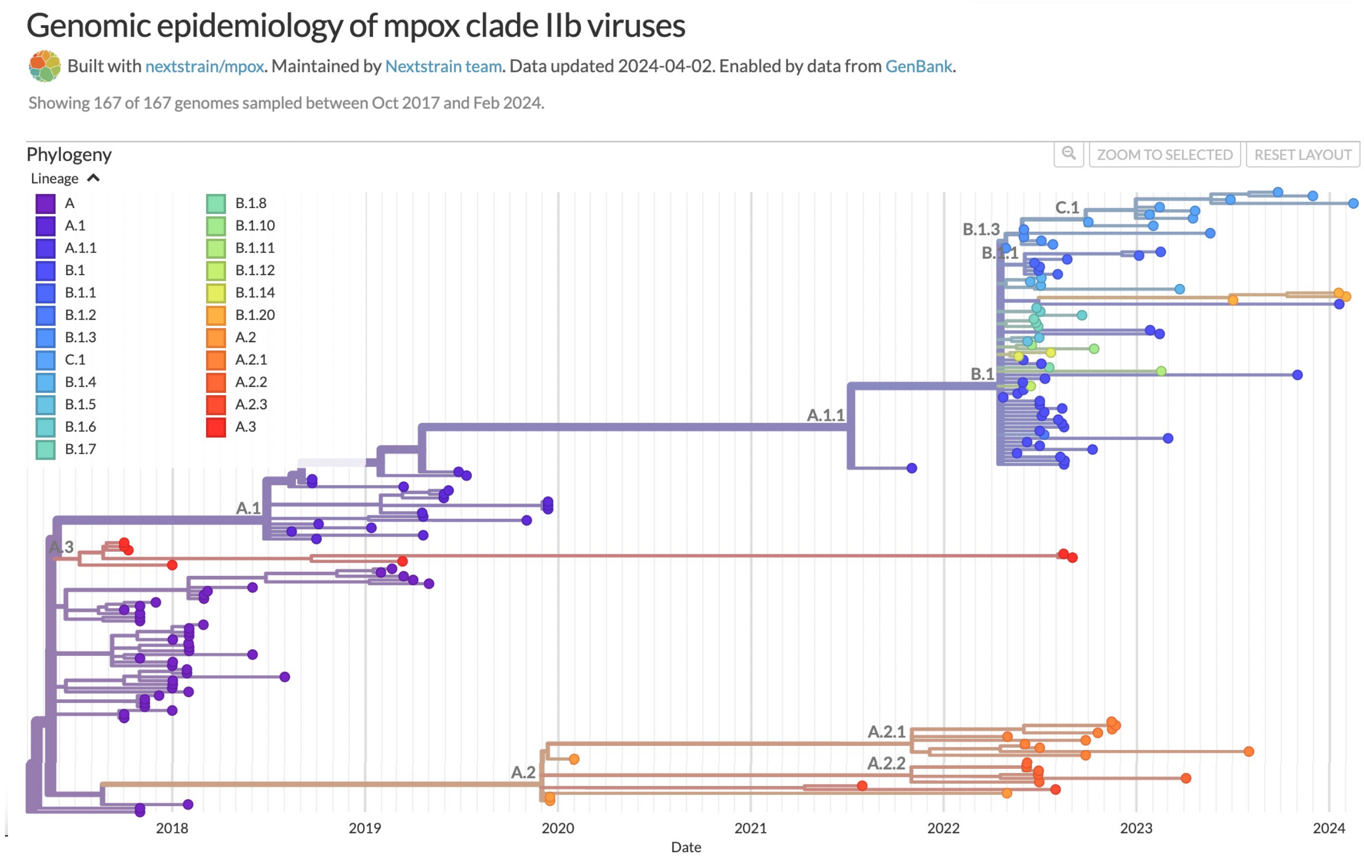


hMPXV1

Maintenance?

Happi et al, PLoS Biol (2022)

2022

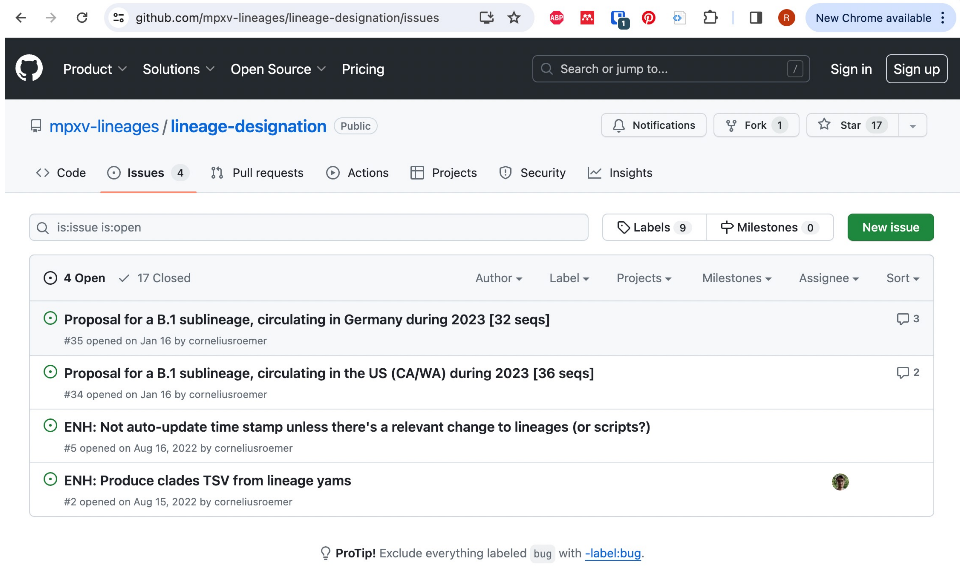


Genomic epidemiology of mpox clade IIb viruses

Built with nextstrain/mpox. Maintained by Nextstrain team. Data updated 2024-04-02. Enabled by data from GenBank.

Showing 167 of 167 genomes sampled between Oct 2017 and Feb 2024.

Nextstrain



github.com/mpxv-lineages/lineage-designation/issues

mpxv-lineages / lineage-designation Public

is:issue is:open

4 Open ✓ 17 Closed

Proposal for a B.1 sublineage, circulating in Germany during 2023 [32 seqs] 3
#35 opened on Jan 16 by corneliusroemer

Proposal for a B.1 sublineage, circulating in the US (CA/WA) during 2023 [36 seqs] 2
#34 opened on Jan 16 by corneliusroemer

ENH: Not auto-update time stamp unless there's a relevant change to lineages (or scripts?)
#5 opened on Aug 16, 2022 by corneliusroemer

ENH: Produce clades TSV from lineage yams
#2 opened on Aug 15, 2022 by corneliusroemer

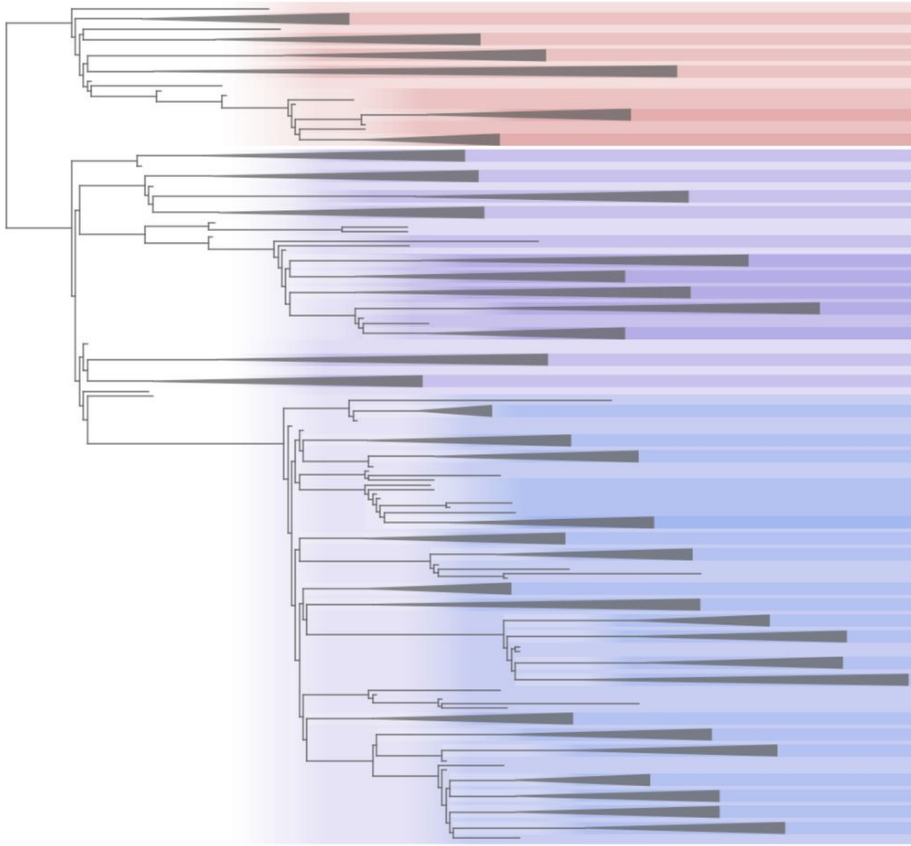
ProTip! Exclude everything labeled bug with -label:bug.

Definitions

Definition of the Pango Lineage system

Pango lineages provide a fine-scale, hierarchical partition of the phylogenetic tree(s)

Pango lineages are designated for the purpose of aiding genomic epidemiology

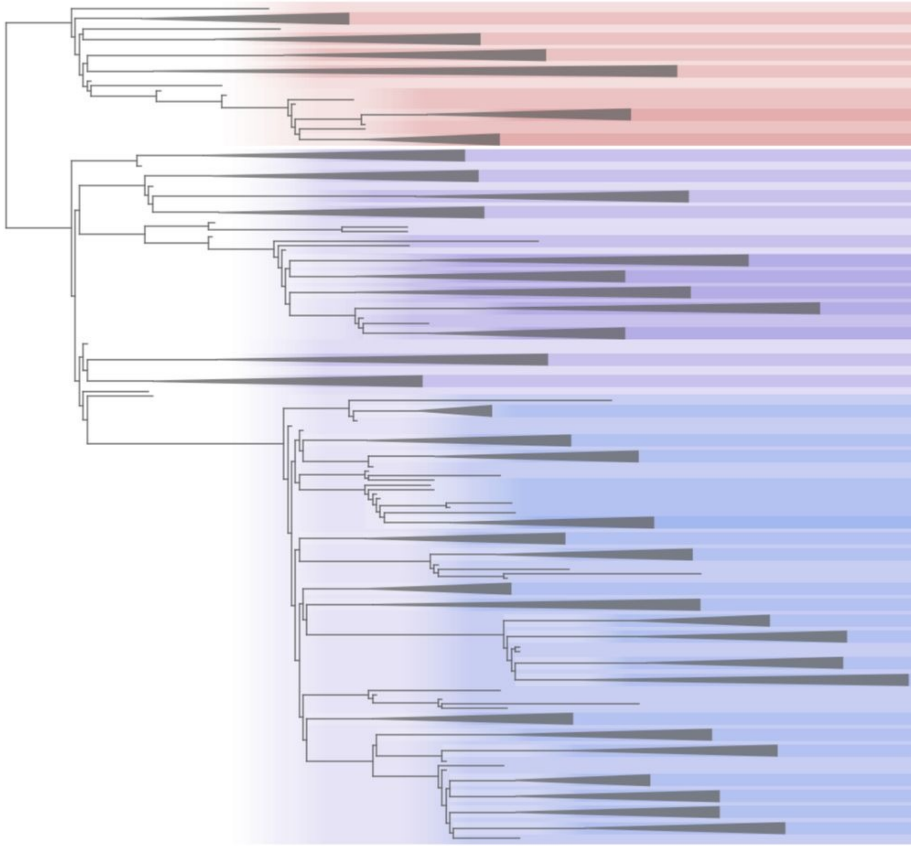


Definition of the Pango Lineage system

Pango lineages provide a fine-scale, hierarchical partition of the phylogenetic tree(s)

Pango lineages are designated for the purpose of aiding genomic epidemiology

There is no requirement for an “even” distribution of lineages



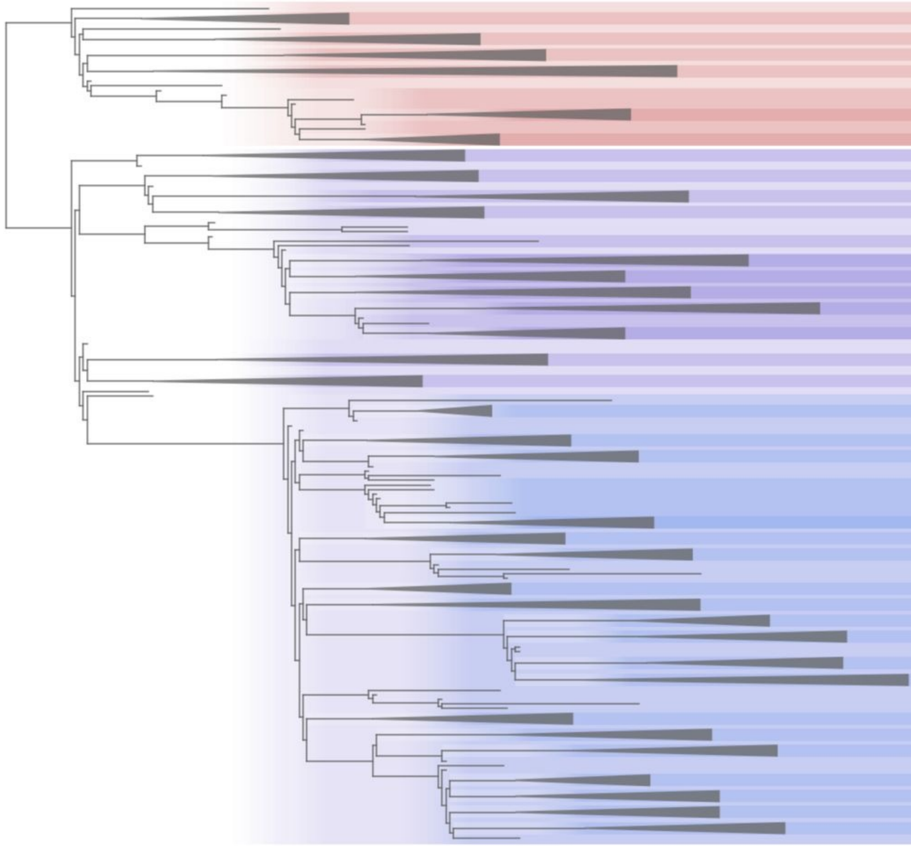
Definition of the Pango Lineage system

Pango lineages provide a fine-scale, hierarchical partition of the phylogenetic tree(s)

Pango lineages are designated for the purpose of aiding genomic epidemiology

There is no requirement for an “even” distribution of lineages

It is not necessary to designate everything



Definition of the Pango Lineage system

Each lineage is given a unique alphanumeric code that contains partial, but not complete, information about the phylogenetic history of that lineage

Definition of the Pango Lineage system

Each lineage is given a unique alphanumeric code that contains partial, but not complete, information about the phylogenetic history of that lineage

e.g.

XBB.1.5

XBB is a recombinant of BJ.1 (BA.2.10.1.1) and BA.2.75 (BA.2.75.3.1.1.1)

Definition of the Pango Lineage system

Each lineage is given a unique alphanumeric code that contains partial, but not complete, information about the phylogenetic history of that lineage

e.g.

XBB.1.5

XBB is a recombinant of BJ.1 (BA.2.10.1.1) and BA.2.75 (BA.2.75.3.1.1.1)

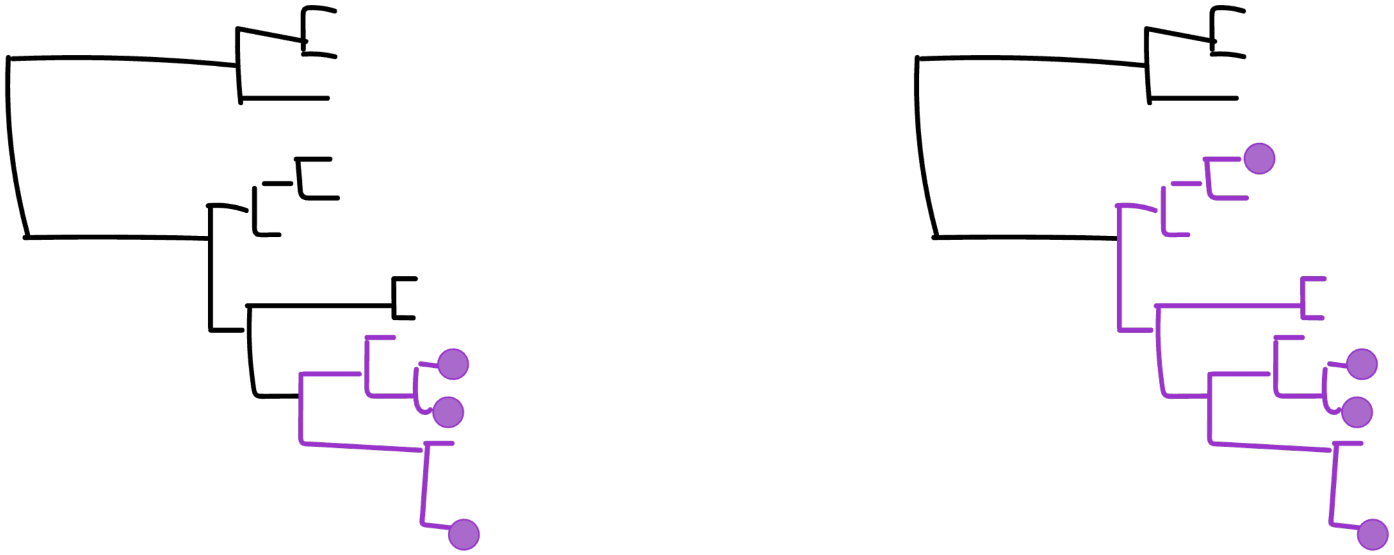
The lineage name should not encode any other features

Definition of the Pango Lineage system

A lineage is entirely defined by the set of designated sequences which represent that lineage

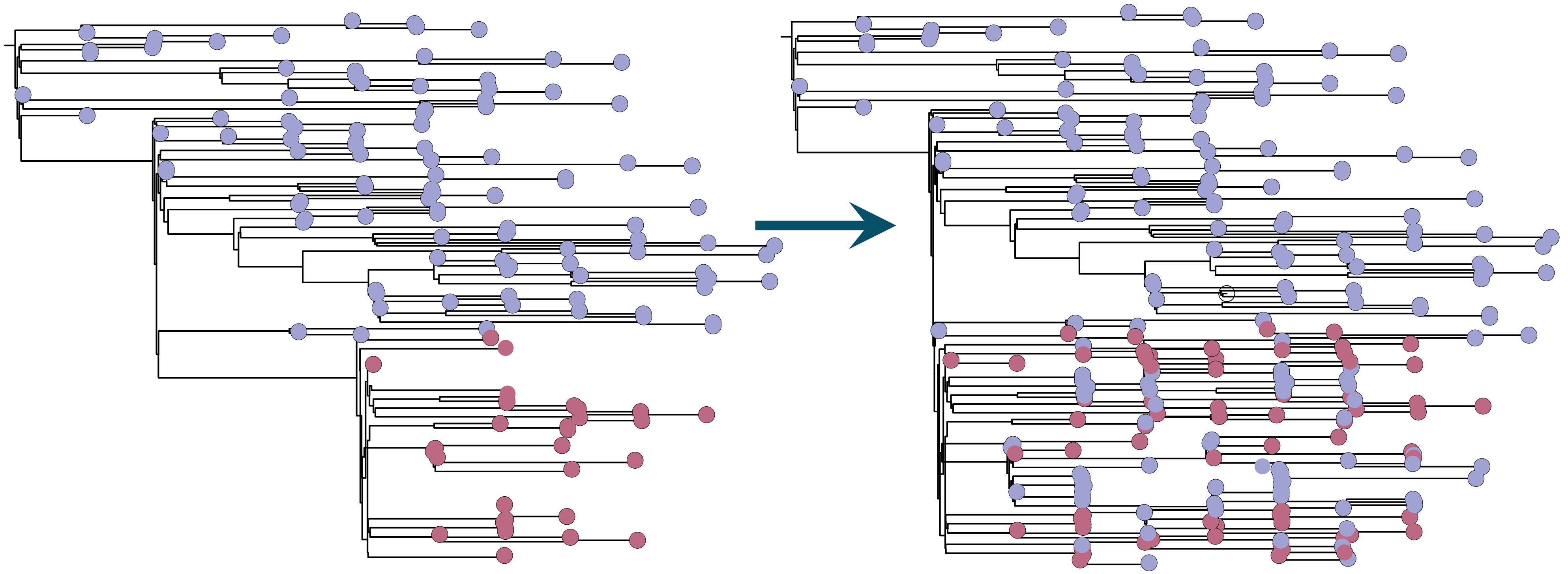
Definition of the Pango Lineage system

A lineage is entirely defined by the set of designated sequences which represent that lineage,



i.e. it is the set of sequences which belong to the minimal clade containing the designated sequences

Lineage definitions may need to change



Data available at time 1

Data available at time 2

Criteria for New Lineage (Original)

A new lineage must meet all of the following criteria:

(a) it exhibits one or more shared nucleotide differences from the ancestral lineage

(b) it comprises at least five genomes with >95% of the genome sequenced

(c) genomes within the lineage exhibit at least one shared nucleotide change among them

(d) a bootstrap value >70% for the lineage-defining node

(e) represent emergence from an ancestral lineage into another geographically distinct population

Criteria for New Lineage (Generalized)

A new lineage must meet all of the following criteria:

(a) it is genetically distinct from the ancestral lineage

(b) there are sufficient high-quality genomes to evaluate

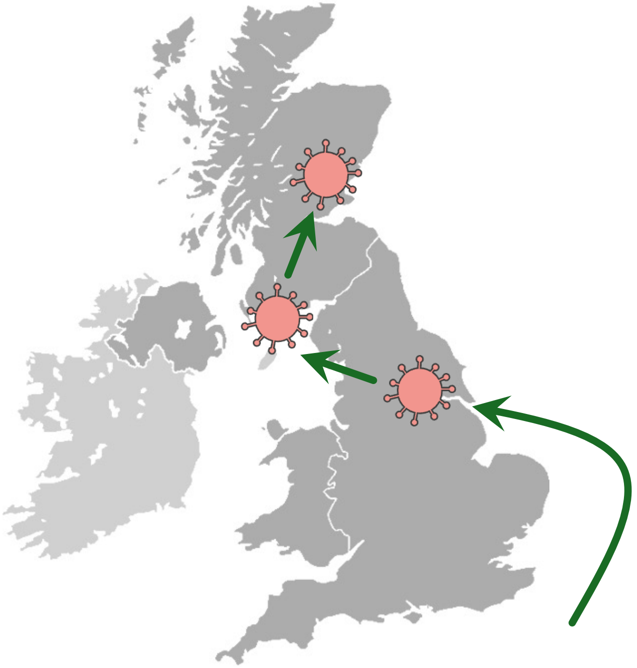
(c) there is evidence of onward transmission in the community

~~(d) the lineage defining node must be well supported~~

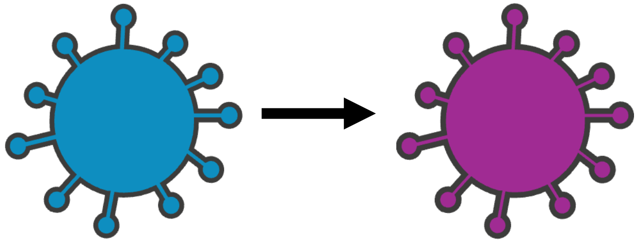
(e) represent an epidemiologically-relevant event

Epidemiological events

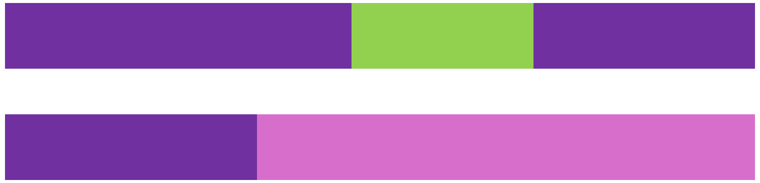
Intro into new region and onward transmission



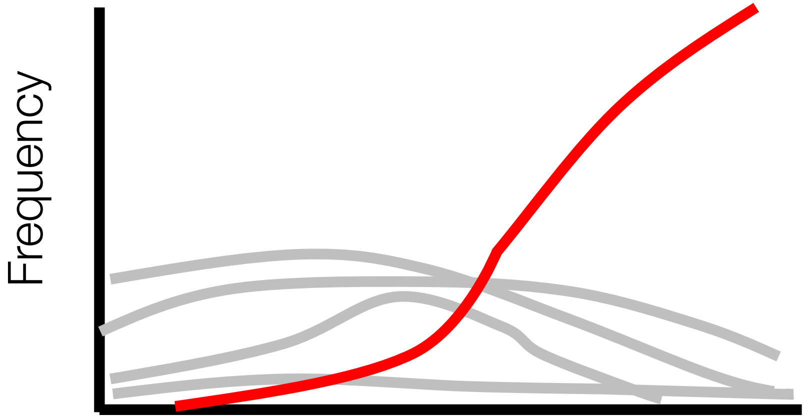
Change in phenotype



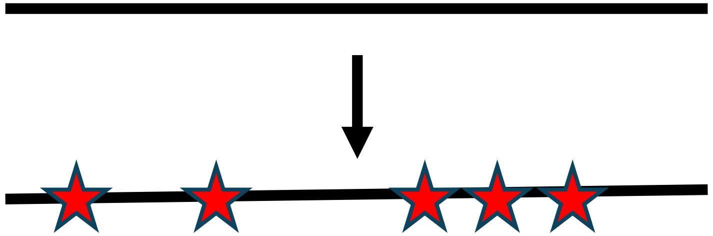
A recombination or reassortment event



Rapid growth compared to other lineages

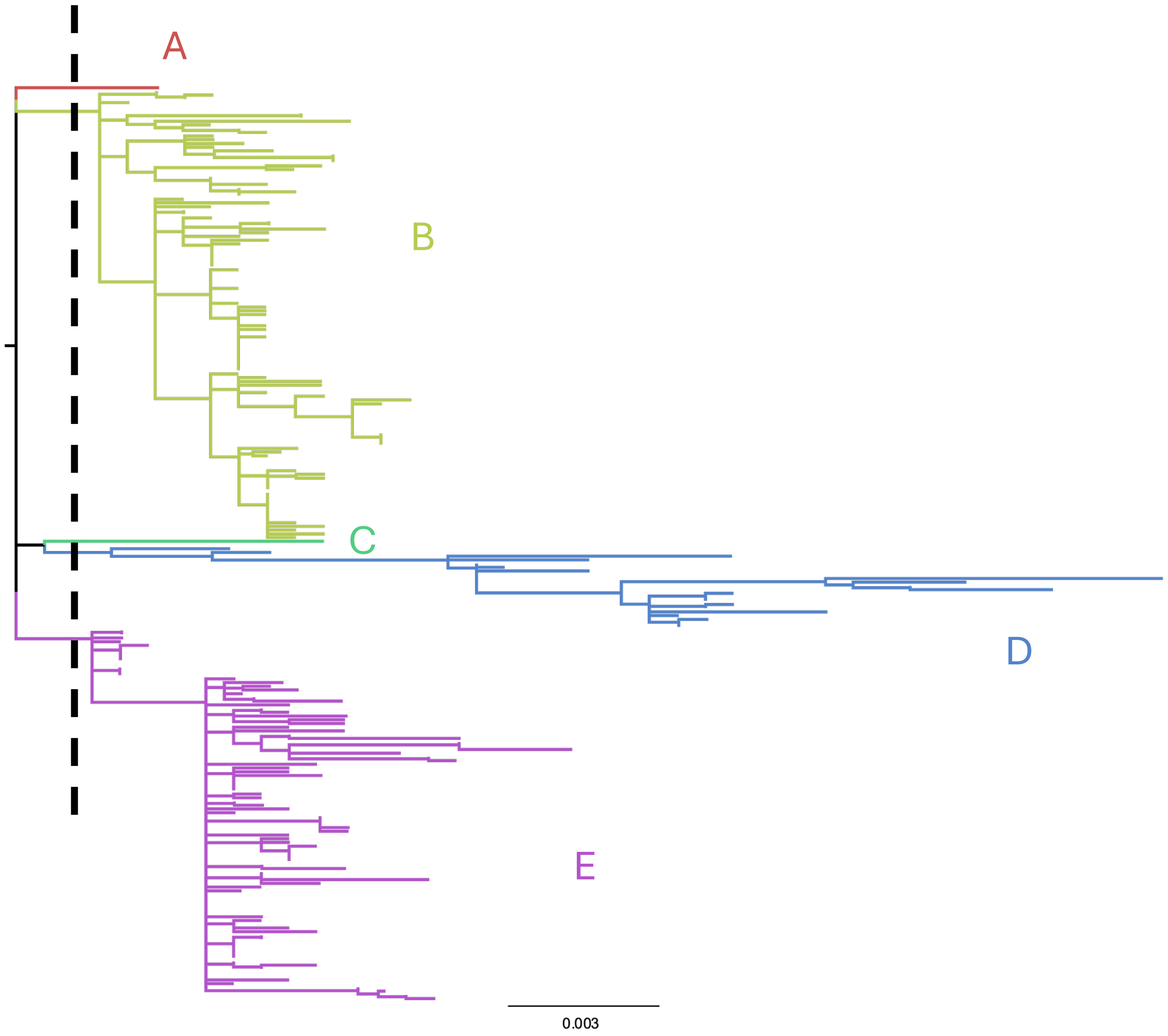


Set of interesting mutations



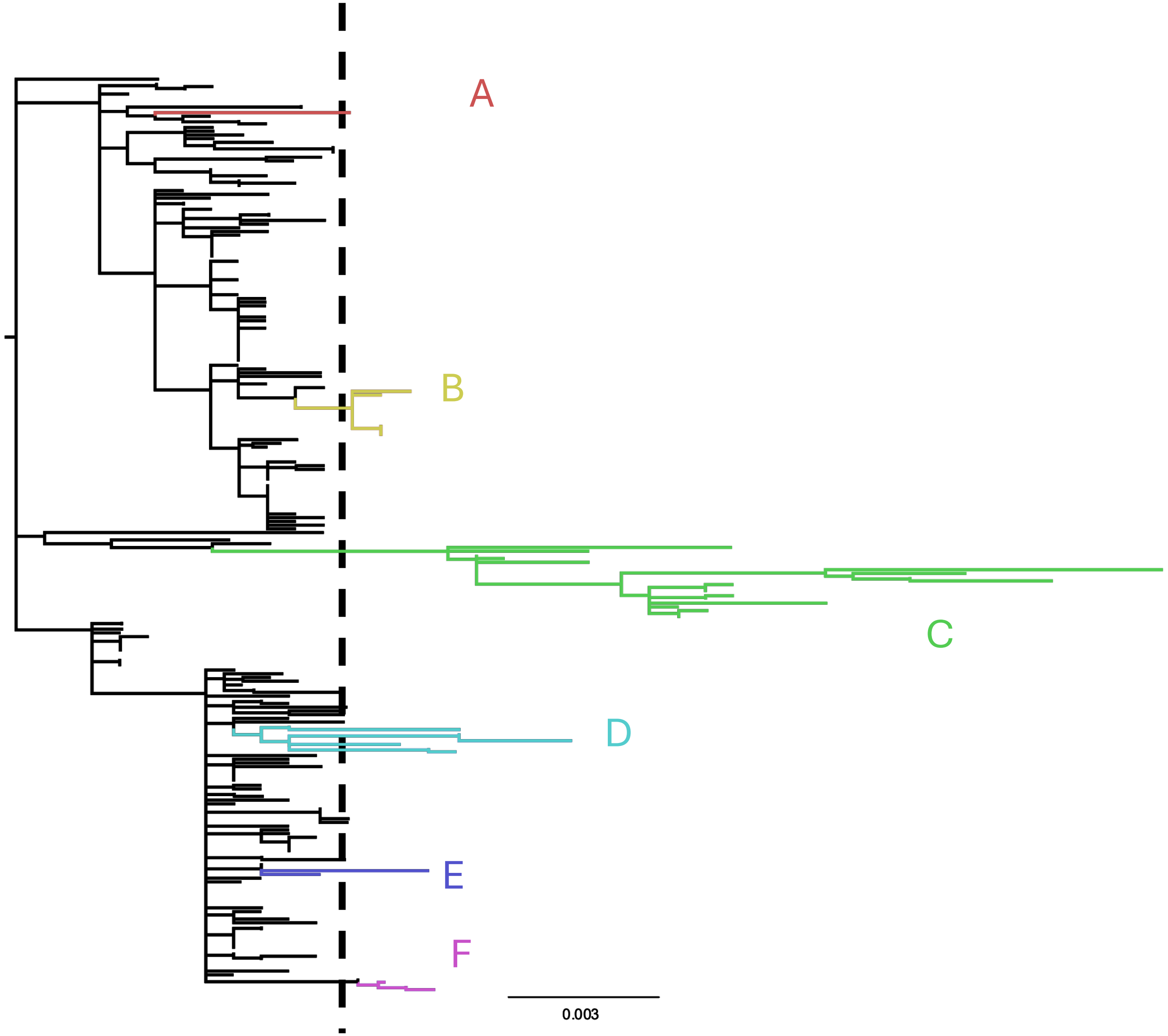
Starting

Set an initial time point. At this cut-off, each phylogenetically distinct cluster of sequences should each be given an alphabetic lineage name A, B, …



Starting

Set an initial time point. At this cut-off, each phylogenetically distinct cluster of sequences should each be given an alphabetic lineage name A, B, …



Aliases

To maintain human-readability, we recommend using aliases after 2 or 3 sub-levels

e.g. B.1.1.1 = C.1

or B.1.1.1.1 = C.1

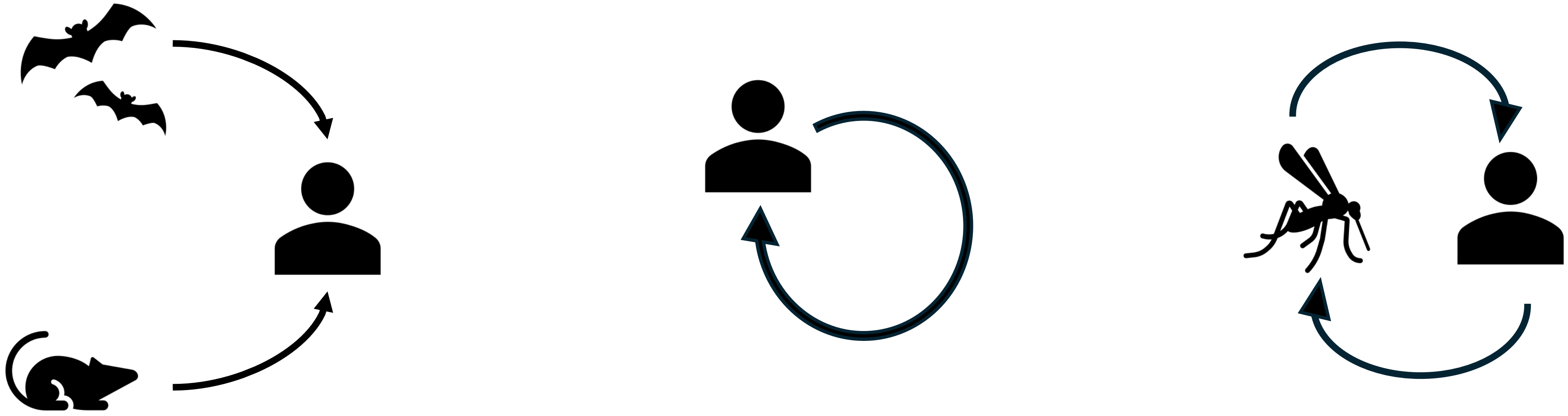
Recombination

Recombination is an epidemiologically relevant event

Given that there is not a single ancestral lineage, a new lineage is given the next available X-prefixed lineage name

Lineages are defined after a host-jump

A Pango lineage nomenclature should be (at least) specific to a single host-jump.



Repeated zoonoses

Each new outbreak qualifies for a new Pango lineage system

e.g. the next MPXV spillover into humans which causes a big enough outbreak would generate a nomenclature for hMPXV2

Reassorting viruses

The nomenclature should be based on the phylogeny for a single segment. If there is a strong reason to do so, a second nomenclature could be created for the phylogeny corresponding to a second segment.

e.g. flu could have a nomenclature for just HA or also NA

Reassorting viruses

The nomenclature should be based on the phylogeny for a single segment. If there is a strong reason to do so, a second nomenclature could be created for the phylogeny corresponding to a second segment.

e.g. flu could have a nomenclature for just HA or also NA

Reassortment events of other segments would qualify as an epidemiological event leading to a new lineage designation. Reassortment within the representative segment would result in an X-prefixed lineage designation

Endemic viruses

There maybe specific cases where we want to track an endemic virus in fine scale, e.g. in response to a new treatment/vaccine

We do not recommend recreating the full history of the phylogeny with lineage designations, but instead identify an initial time cut-off and split the phylogeny of sequences circulating after that time into lineages A, B, C, …

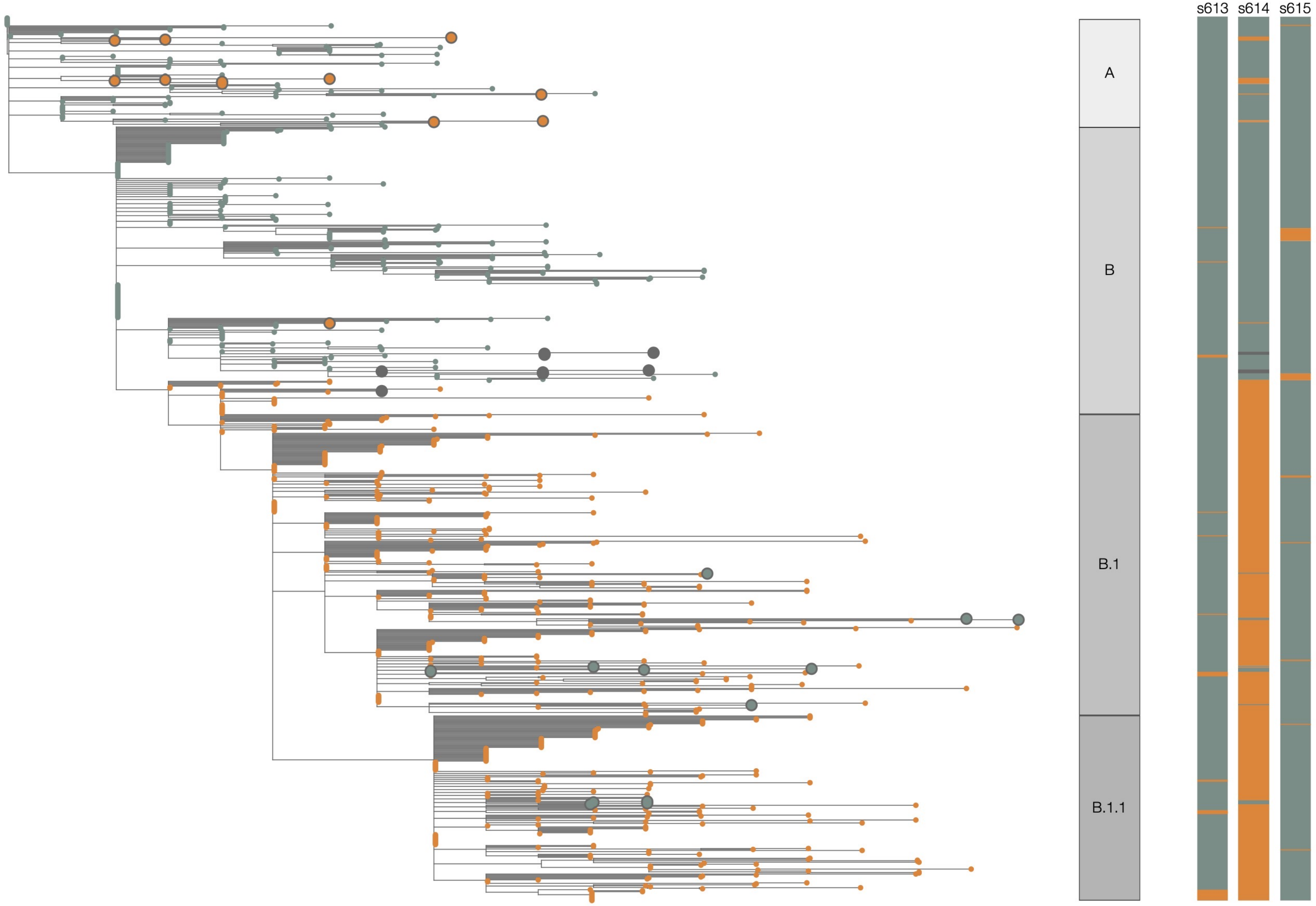
Variants

A variant has a specific signature of mutations

Lineages tracking may help identify potential variants

BUT not all variants are lineages

Homoplasies (e.g. D614G)



Variants

A variant has a specific signature of mutations

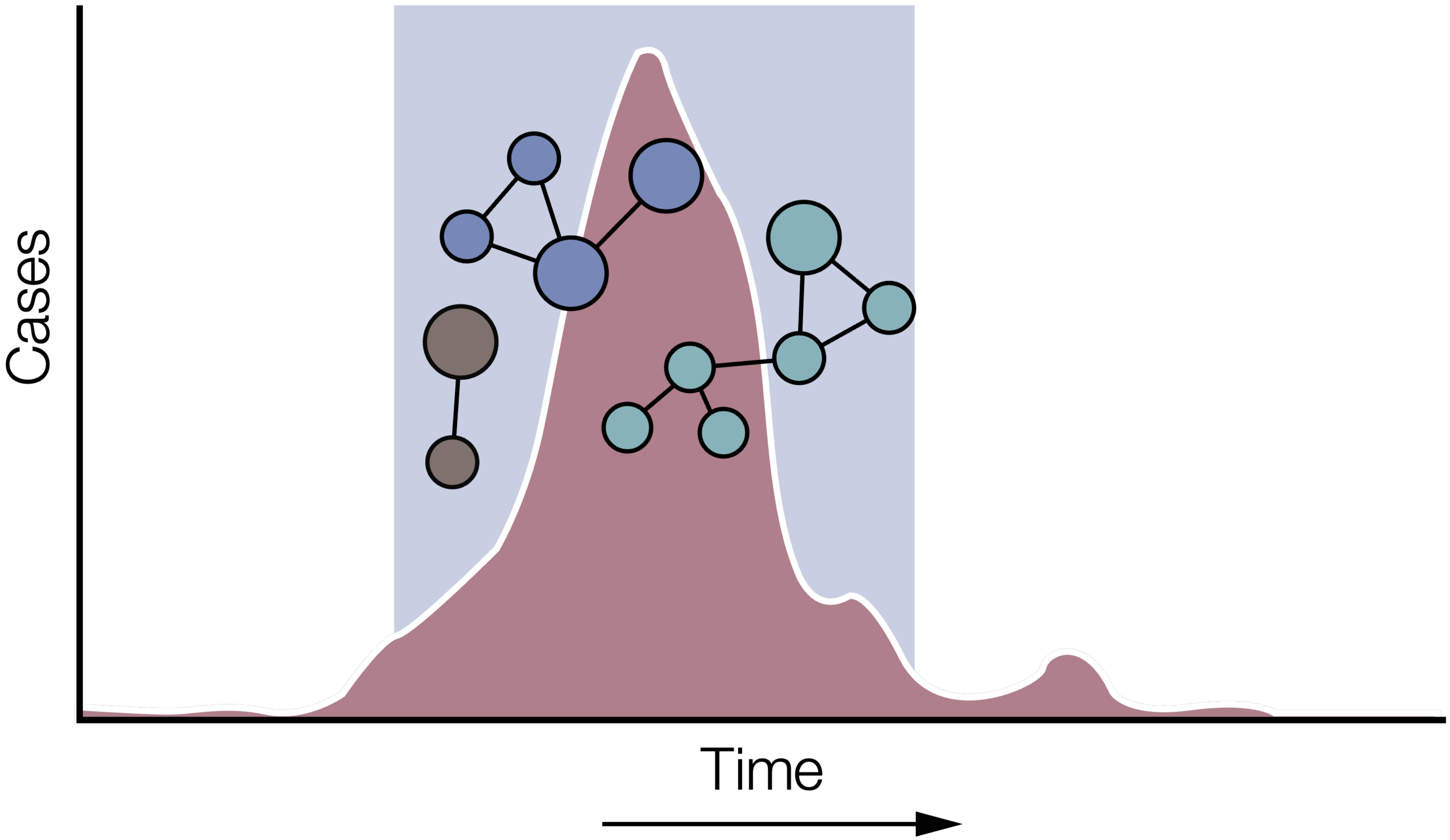
Lineages tracking may help identify potential variants

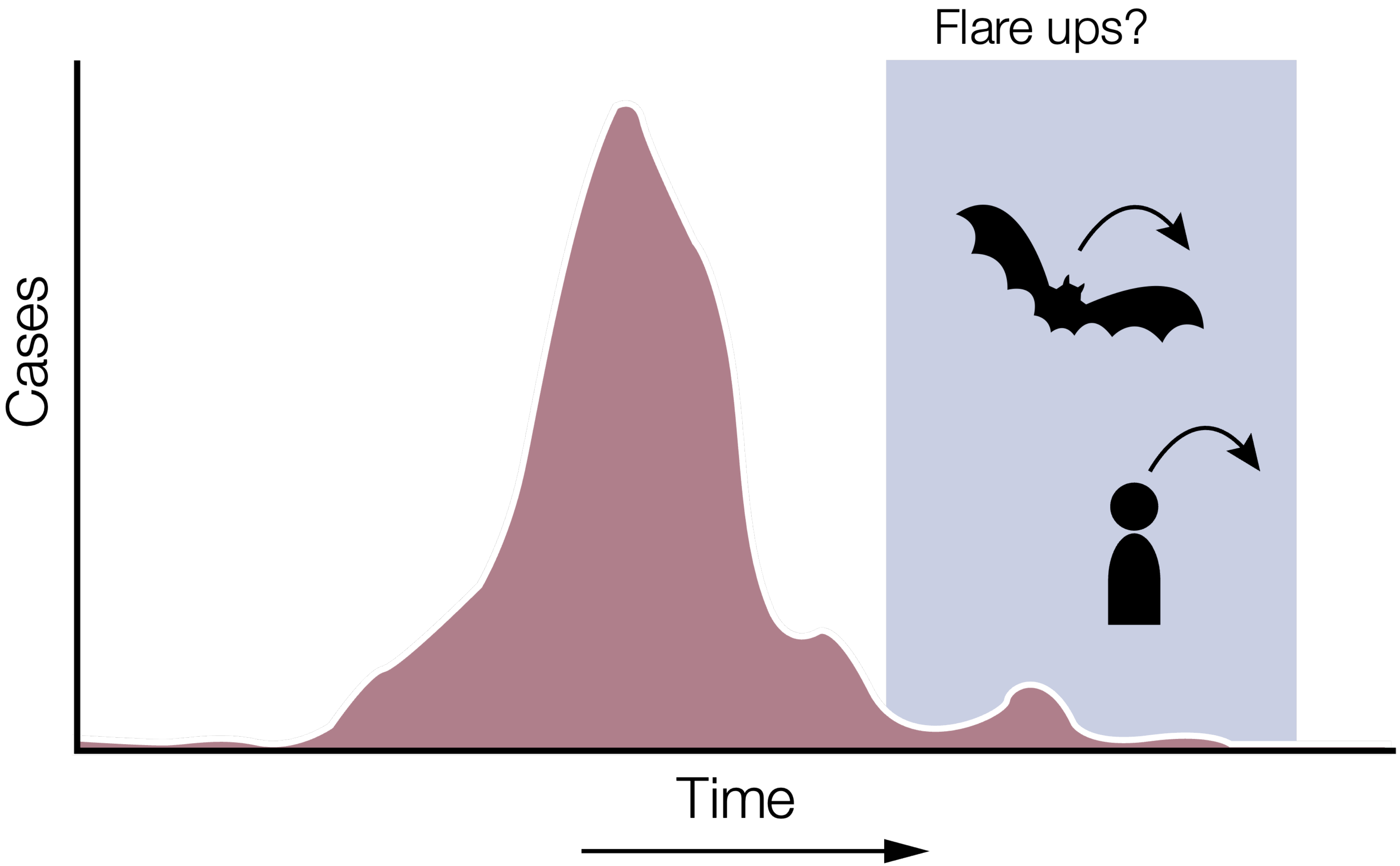
BUT not all variants are lineages

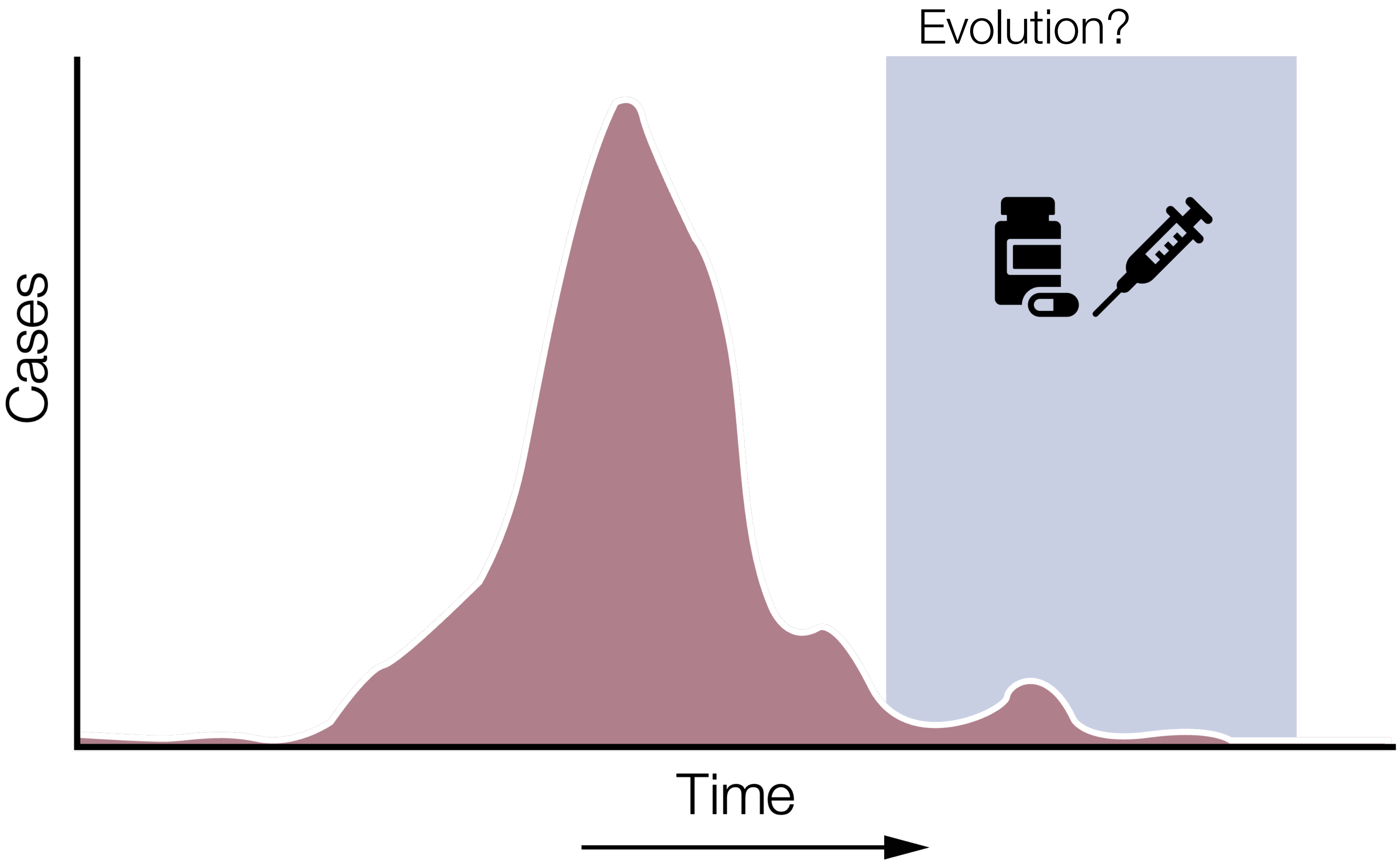
Variant typing needs to be explicit about the defining mutations (ref, alt, ambiguous) e.g. scorpio

Challenges

Modes of transmission
Transmission chains




Flare ups?
Cases
Time


Evolution?
Cases
Time

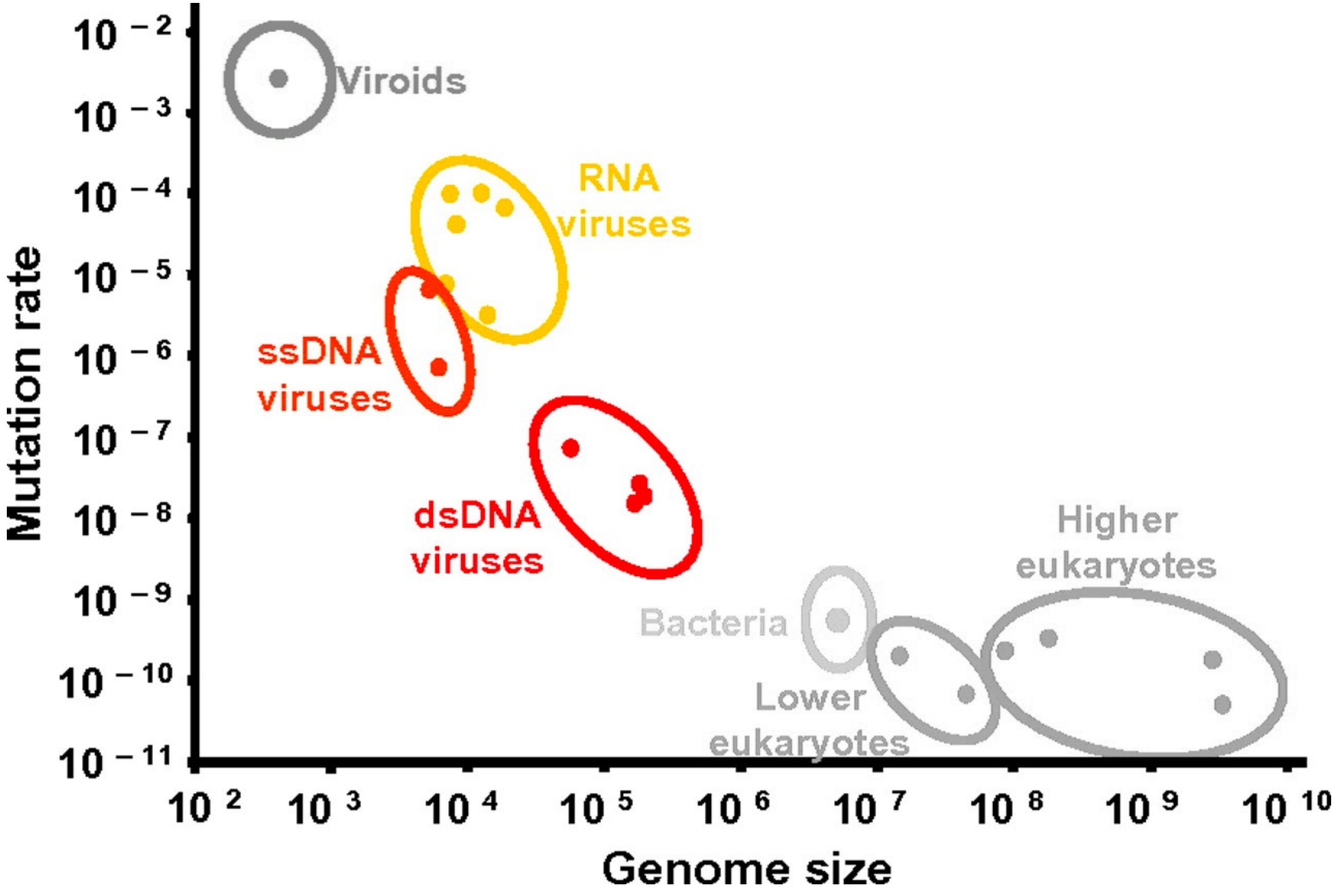
Does the virus have sufficient diversity?

Short generation times

High mutation rates

Small genome size

Not all DNA viruses will accumulate sufficient diversity to track over short time scales



Eddie Holmes, 2009 PNAS

How much recombination/reassortment?

These make phylogenetic methods challenging

What sort of tree?

BEAST

Bayesian Evolutionary Analysis Sampling Trees

UShER

FastTree2

Nextstrain

What sort of tree?

BEAST

Bayesian Evolutionary Analysis Sampling Trees

UShER

FastTree2

Nextstrain

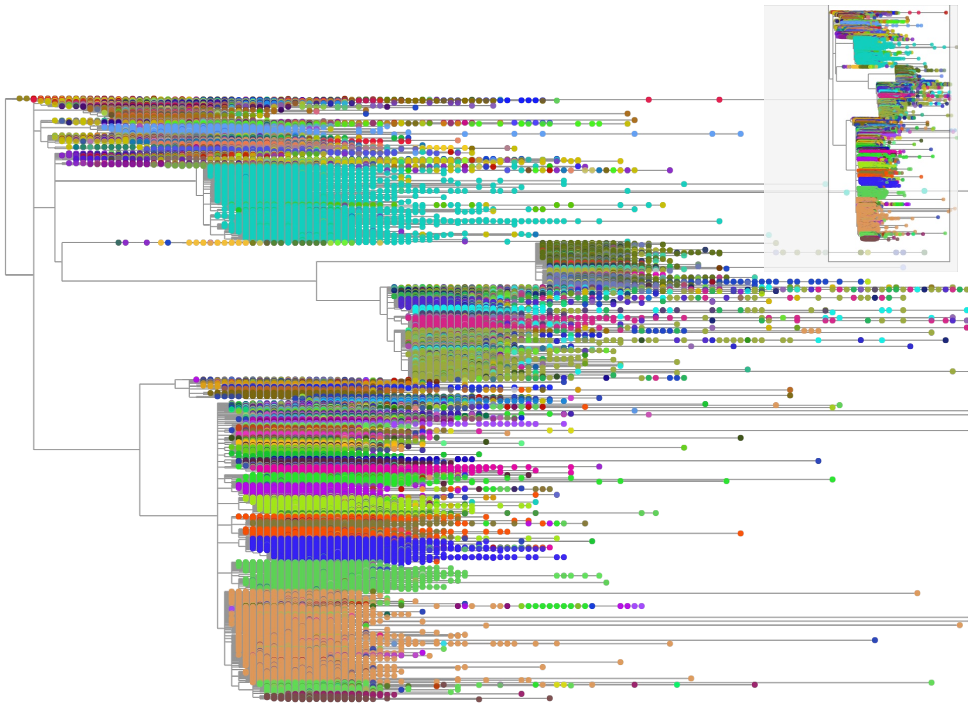
How much virus diversity?

How many sequences?

Can we automate tree construction?

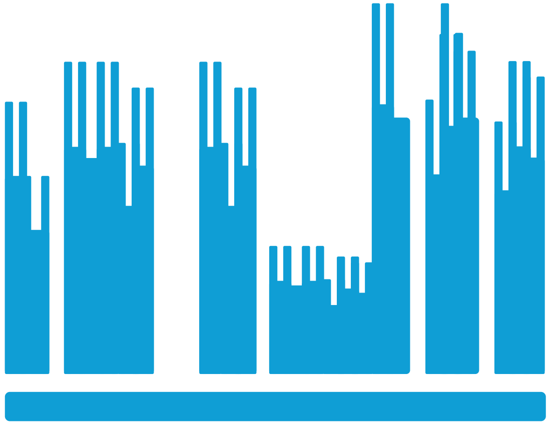
UShER

Nextstrain

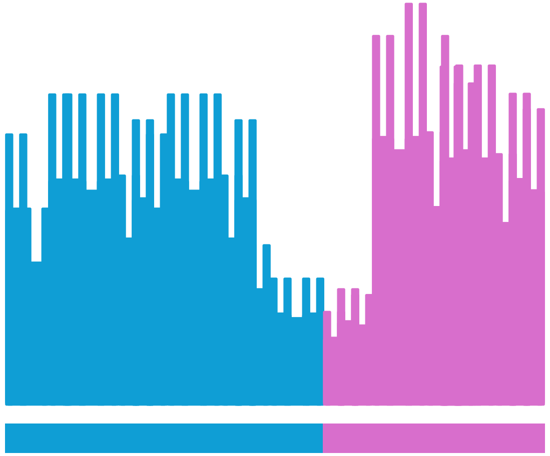


Cov2Tree powered by Taxonium

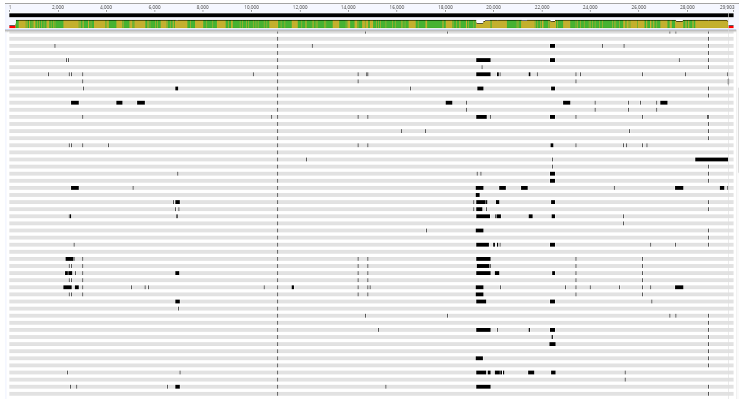
Can we automate tree construction?



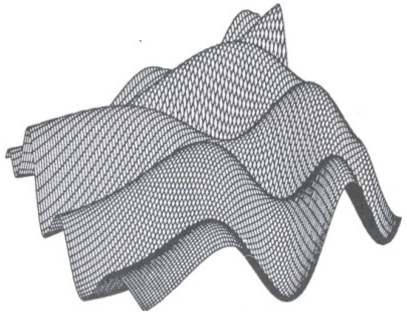
Low-coverage calls to reference



Contamination vs recombination

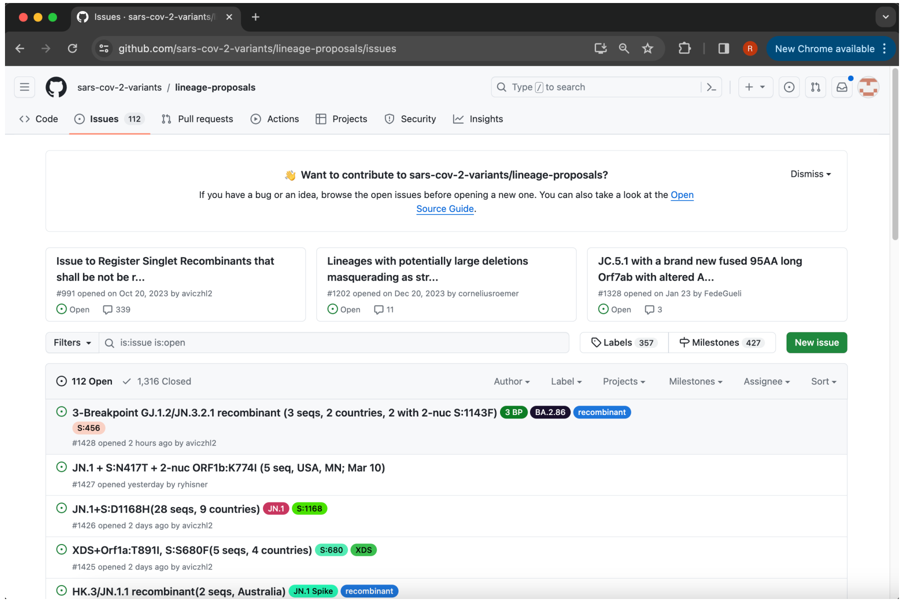


Amplicon dropouts



What is the true tree?

Can we automate lineage designation?



👋 Want to contribute to sars-cov-2-variants/lineage-proposals?

If you have a bug or an idea, browse the open issues before opening a new one. You can also take a look at the Open Source Guide.

Issue to Register Singlet Recombinants that shall be not be r...
#991 opened on Oct 20, 2023 by aviczhl2
Open 339

Lineages with potentially large deletions masquerading as str...
#1202 opened on Dec 20, 2023 by corneliusroemer
Open 11

JC.5.1 with a brand new fused 95AA long Orf7ab with altered A...
#1328 opened on Jan 23 by FedeGueli
Open 3

is:issue is:open

112 Open ✓ 1,316 Closed

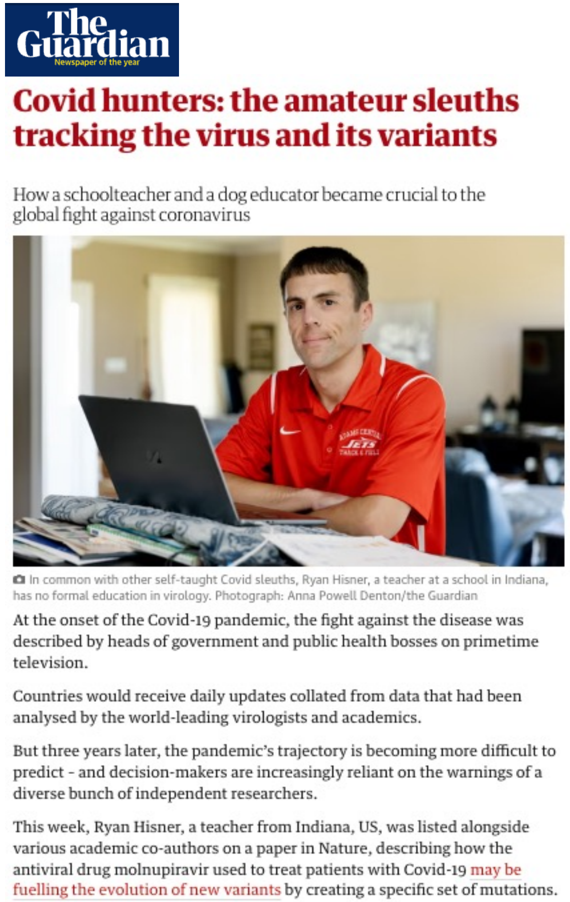
- 3-Breakpoint GJ.1.2/JN.3.2.1 recombinant (3 seqs, 2 countries, 2 with 2-nuc S:1143F) 3 BP BA.2.86 recombinant S:456
  #1428 opened 2 hours ago by aviczhl2
- JN.1 + S:N417T + 2-nuc ORF1b:K774I (5 seq, USA, MN; Mar 10)
  #1427 opened yesterday by ryhisner
- JN.1+S:D1168H(28 seqs, 9 countries) JN.1 S:1168
  #1426 opened 2 days ago by aviczhl2
- XDS+Orf1a:T891I, S:S680F(5 seqs, 4 countries) S:680 XDS
  #1425 opened 2 days ago by aviczhl2
- HK.3/JN.1.1 recombinant(2 seqs, Australia) JN.1 Spike recombinant

Variant Spotter Community

Cornelius Roemer
Federico Gueli
Xu Zou

Ryan Hisner
Yat-sing Ng

...and more


The Guardian
Newspaper of the year

Covid hunters: the amateur sleuths tracking the virus and its variants

How a schoolteacher and a dog educator became crucial to the global fight against coronavirus

In common with other self-taught Covid sleuths, Ryan Hisner, a teacher at a school in Indiana, has no formal education in virology. Photograph: Anna Powell Denton/the Guardian

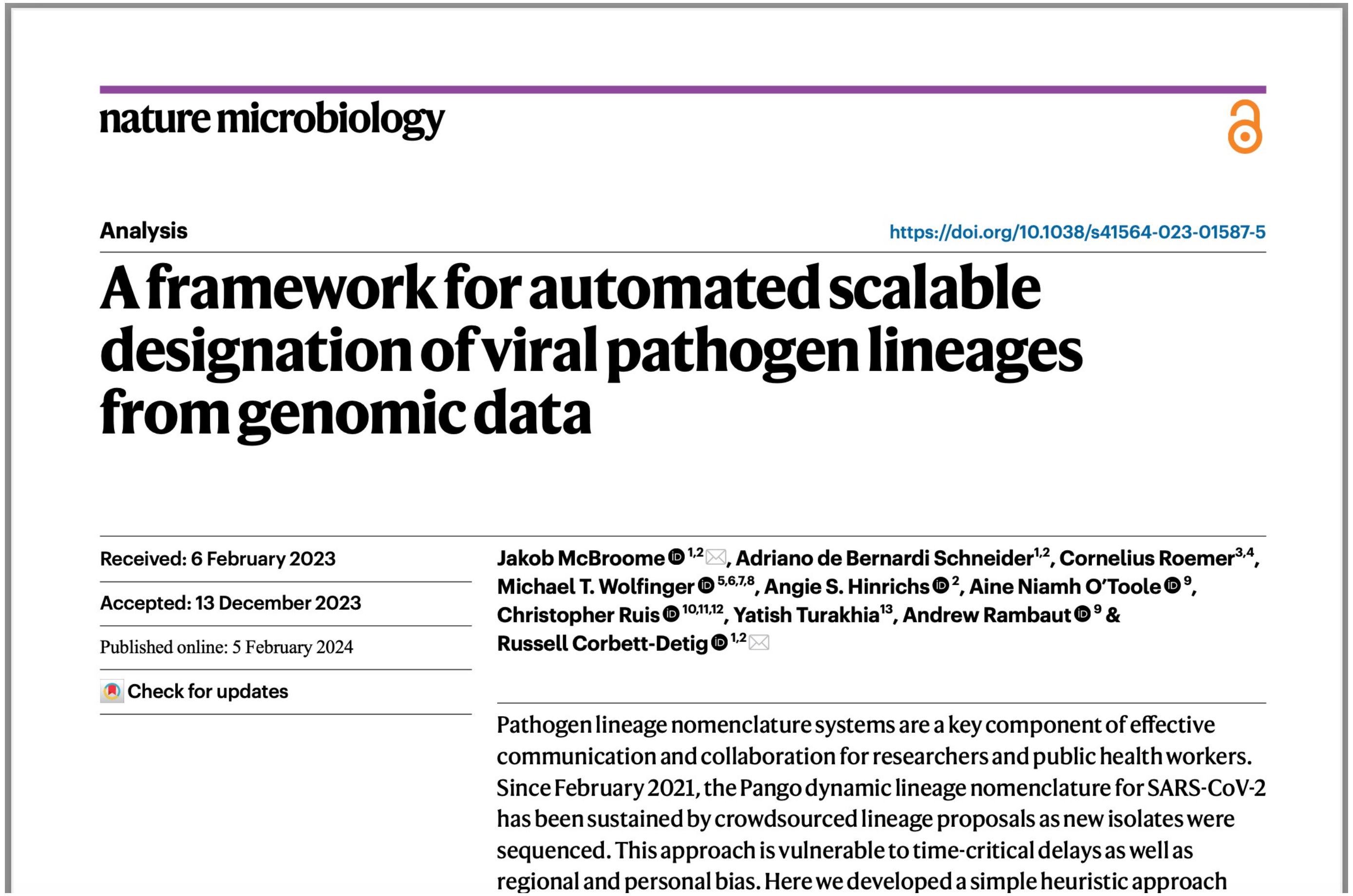
At the onset of the Covid-19 pandemic, the fight against the disease was described by heads of government and public health bosses on primetime television.

Countries would receive daily updates collated from data that had been analysed by the world-leading virologists and academics.

But three years later, the pandemic's trajectory is becoming more difficult to predict – and decision-makers are increasingly reliant on the warnings of a diverse bunch of independent researchers.

This week, Ryan Hisner, a teacher from Indiana, US, was listed alongside various academic co-authors on a paper in Nature, describing how the antiviral drug molnupiravir used to treat patients with Covid-19 may be fuelling the evolution of new variants by creating a specific set of mutations.

Can we automate lineage designation?



nature microbiology

Analysis

https://doi.org/10.1038/s41564-023-01587-5

A framework for automated scalable designation of viral pathogen lineages from genomic data

Received: 6 February 2023

Accepted: 13 December 2023

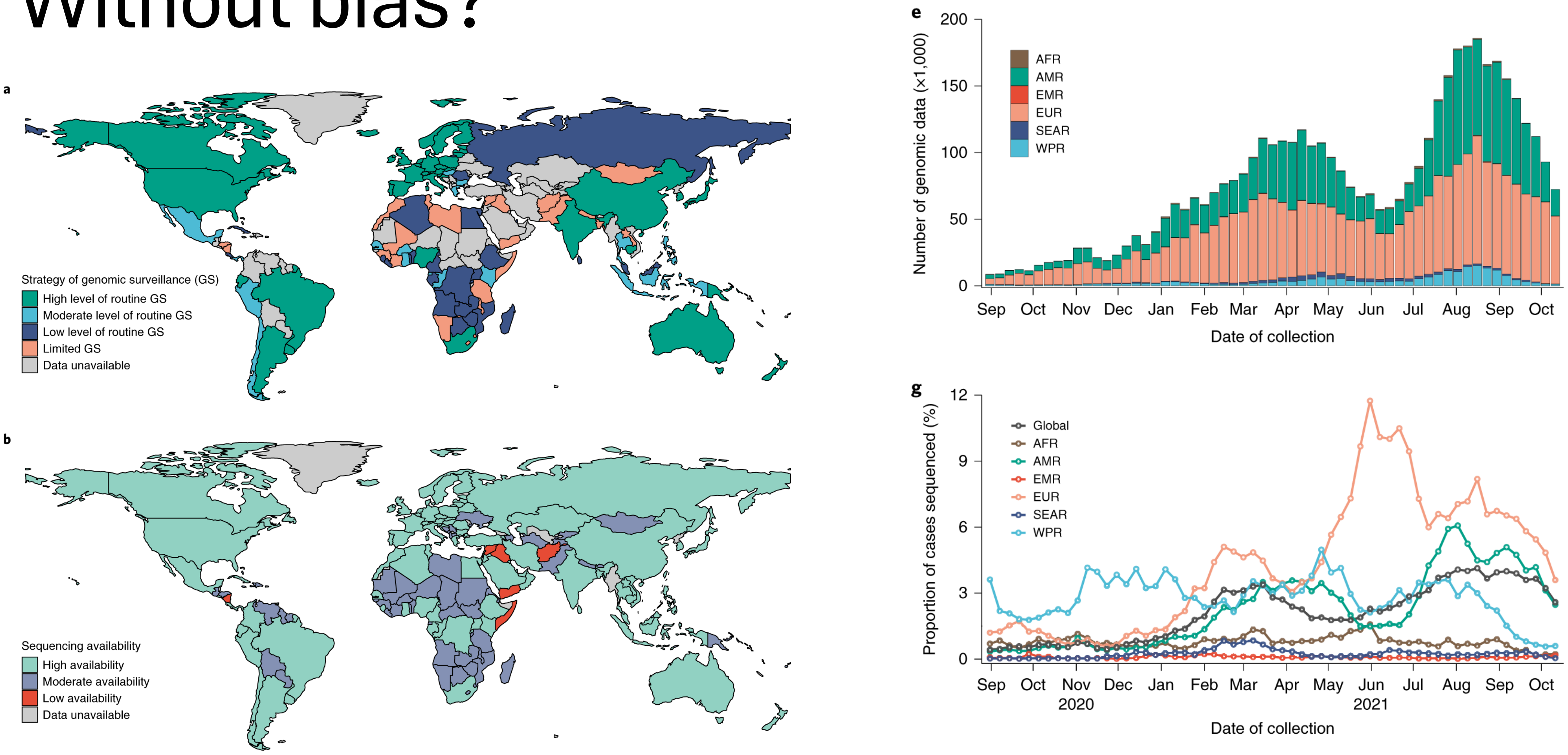
Published online: 5 February 2024

Check for updates

Jakob McBroome[1,2], Adriano de Bernardi Schneider[1,2], Cornelius Roemer[3,4], Michael T. Wolfinger[5,6,7,8], Angie S. Hinrichs[2], Aine Niamh O'Toole[9], Christopher Ruis[10,11,12], Yatish Turakhia[13], Andrew Rambaut[9] & Russell Corbett-Detig[1,2]

Pathogen lineage nomenclature systems are a key component of effective communication and collaboration for researchers and public health workers. Since February 2021, the Pango dynamic lineage nomenclature for SARS-CoV-2 has been sustained by crowdsourced lineage proposals as new isolates were sequenced. This approach is vulnerable to time-critical delays as well as regional and personal bias. Here we developed a simple heuristic approach

Without bias?

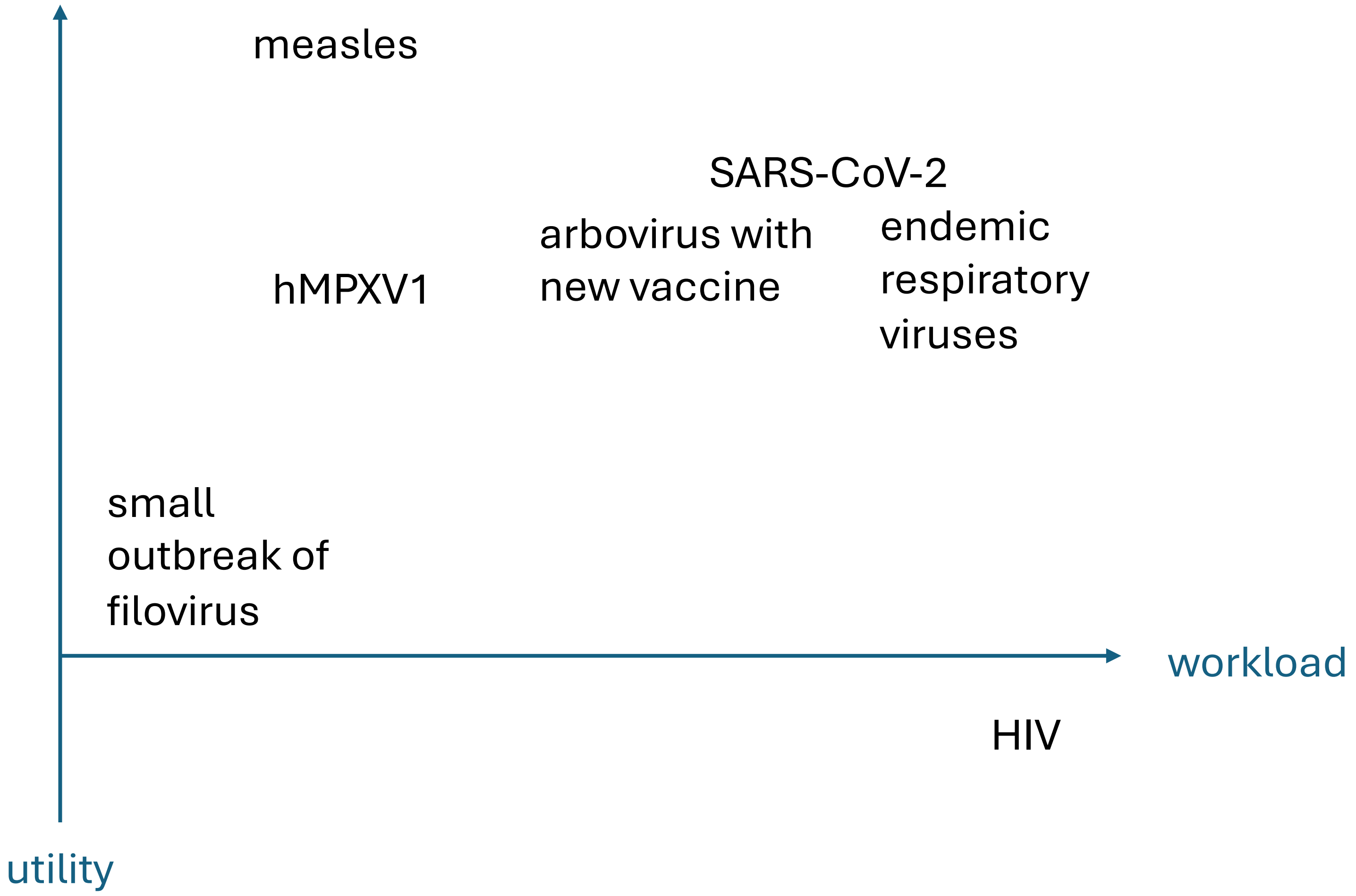


Chen, Z., Azman, A.S., Chen, X. *et al.* Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat Genet* **54**, 499–507 (2022).

We can’t avoid manual input…

We can’t avoid manual input…

…so it needs to be worth it


measles
SARS-CoV-2
arbovirus with new vaccine
endemic respiratory viruses
hMPXV1
small outbreak of filovirus
workload
HIV
utility

Can we make a Pango Lineage nomenclature (is the viral evolution on the scale required)?
Would “a tree” be sufficient for this (eg small outbreak transmission chains)?
What is the utility gained by using Pango Lineages (are we searching for new variants, global transmission chains, immune escape)?

Thank you!

University of Edinburgh

Andrew Rambaut
Áine O’Toole
Verity Hill
Danny Maloney
JT McCrone
Ben Jackson
Emily Scher
Shawn Yu
Corey Ansley
Ifeanyi Omah
Kate Duggan
Zoe Vance

Pango Network

Oliver Pybus
Chris Ruis
Tom Peacock
Stephen Attwood
Angie Hinrichs
Cornelius Roemer
Eddie Holmes
Louis du Plessis
et al

Variant Spotters Community

Cornelius Roemer
Tom Peacock
Federico Gueli
Xu Zou
Ryan Hisner
Yat-sing Ng
et al

THE UNIVERSITY of EDINBURGH

FAST — GRANTS

OXFORD MARTIN SCHOOL

wellcome