VIRUS EVOLUTION

Virus Evolution, 2024, **10(1)**, 1–11
DOI: https://doi.org/10.1093/ve/vead085
Advance Access Publication 11 January 2024
Research Article

SARS-CoV-2 lineage assignments using phylogenetic placement/UShER are superior to pangoLEARN machine-learning method

Adriano de Bernardi Schneider,[1,2,‡,†,*] Michelle Su,[3,†] Angie S. Hinrichs,[1,§] Jade Wang,[3] Helly Amin,[3] John Bell,[4] Debra A. Wadford,[4] Áine O'Toole,[5,**] Emily Scher,[5] Marc D. Perry,[1] Yatish Turakhia,[6,††] Nicola De Maio,[7,‡‡] Scott Hughes,[3] and Russ Corbett-Detig[1,2]

[1]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA, [2]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA, [3]Department of Health and Mental Hygiene, New York City Public Health Laboratory, New York, NY 10016, USA, [4]California Department of Public Health (CDPH), VRDL/COVIDNet, Richmond, CA 94804, USA, [5]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK, [6]Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA 92093, USA and [7]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton CB10 1SD, UK
[†]These authors contributed equally to this work
[‡]https://orcid.org/0000-0001-7487-266X
[§]https://orcid.org/0000-0002-1697-1130
[**]https://orcid.org/0000-0001-8083-474X
[††]https://orcid.org/0000-0001-5600-2900
[‡‡]https://orcid.org/0000-0002-1776-8564
*Corresponding authors: E-mail: adeberna@ucsc.edu; adrianoidaho@gmail.com

Abstract

With the rapid spread and evolution of SARS-CoV-2, the ability to monitor its transmission and distinguish among viral lineages is critical for pandemic response efforts. The most commonly used software for the lineage assignment of newly isolated SARS-CoV-2 genomes is pangolin, which offers two methods of assignment, pangoLEARN and pUShER. PangoLEARN rapidly assigns lineages using a machine-learning algorithm, while pUShER performs a phylogenetic placement to identify the lineage corresponding to a newly sequenced genome. In a preliminary study, we observed that pangoLEARN (decision tree model), while substantially faster than pUShER, offered less consistency across different versions of pangolin v3. Here, we expand upon this analysis to include v3 and v4 of pangolin, which moved the default algorithm for lineage assignment from pangoLEARN in v3 to pUShER in v4, and perform a thorough analysis confirming that pUShER is not only more stable across versions but also more accurate. Our findings suggest that future lineage assignment algorithms for various pathogens should consider the value of phylogenetic placement.

Keywords: Phylogenetics; Bioinformatics; COVID-19; variants.

Introduction

Determining the genetic relationships between virus strains is key to SARS-CoV-2 surveillance and outbreak investigation. Lineage nomenclature systems have been a constant topic of discussion in the specialized literature with no clearly established nomenclature system for the subclassification of infectious agent lineages or subtypes and significant variability between each pathogen-specific research community (de Bernardi Schneider et al. 2019). Early attempts at microbial lineage classification included the use of technologies such as pulsed-field gel electrophoresis (PFGE) (Khambaty, Bennett, and Shah 1994; Miranda et al. 1996; Thong, Puthucheary, and Pang 1998; Jang et al. 2005; Sandt et al. 2006), but whole-genome sequencing (WGS) has revolutionized the field and given more discriminatory power as well as the ability to characterize additional traits such as strain antimicrobial resistance markers, virulence genes, and plasmid content in one assay (Gilmour et al. 2010; Den Bakker et al. 2014; Jackson et al. 2016; Stucki et al. 2016; Moura et al. 2017; Jajou et al. 2018). WGS has enabled phylogenetic studies to evaluate the relationships of individual sequences and allowed the development of a comprehensive lineage classification system based on genome evolution, a significant improvement over the use of single-gene evolution or phenetics (Durand et al. 2018). Moreover, the use of phylogenetic tools with WGS enables improved epidemiological responses in the field by revealing viral dynamics such as in the 2013–16 Ebola epidemic in West Africa (Dudas et al. 2017). Currently, tools such as Nextstrain (Hadfield et al. 2018) and Nextclade (Aksamentov et al. 2021) are an accessible way to perform phylodynamics and clade assignment for numerous viruses (e.g., West Nile virus, influenza virus, and mpox virus).

SARS-CoV-2 was first reported in December 2019, and by March 2020 it was classified as a pandemic by the World Health Organization (WHO). The rapid spread and lack of test availability in the beginning of the pandemic meant that traditional methods were inadequate to describe the scale of the pandemic. Whole-genome sequencing of SARS-CoV-2 and the subsequent creation of Pango lineages (Rambaut et al. 2020, 2021) have been central in aiding health officials to trace the spread of the virus locally and globally, and identifying differences among viral lineages (O'Toole et al. 2021). Currently, the most commonly used tool for the lineage assignment of newly isolated SARS-CoV-2 genomes is Phylogenetic Assignment of Named Global Outbreak Lineages (pangolin), which offers parsimony-based lineage assignment using pangolin Ultrafast Sample Placement on Existing tRees (pUShER) (default on v4) and pangoLEARN (default on v3) lineage designation modes (CovLineages; Turakhia et al. 2021; Scher, O'Toole, and Rambaut 2022). PangoLEARN aims for a rapid assignment of lineages using a decision tree algorithm. pUShER performs a phylogenetic placement using a maximum parsimony approach to identify the lineage corresponding to a newly sequenced genome. PangoLEARN is substantially faster than pUShER. However, because the Pango lineage nomenclature system is phylogenetic (Rambaut et al. 2020), it is possible that pUShER is more accurate and stable in lineage assignments across subsequent releases. Given the epidemiological importance of assigning strains the correct lineages, we sought to evaluate the consistency and accuracy of the two main methods of lineage assignment currently available (Zhang, Wu, and Zhang 2020).

Despite high overall concordance between pangoLEARN and pUShER lineage assignments, pangoLEARN can be unreliable when new lineages are designated, leading to sequences that must be reassigned in a later software version despite high-genome coverage/quality. In addition, greater single-nucleotide polymorphism (SNP) distances are found between samples that are assigned the same lineage by pangoLEARN (decision tree model but not random forest model). Also, more serious constellations of reoccurring phylogenetically independent origin (Scorpio) lineage call overrides are found in the pangoLEARN analyses. Therefore, we could conclude that pUShER is a more stable and accurate method to assign pangolin lineages to SARS-CoV-2 sequences. More generally, phylogenetic placement is an appealing method for lineage assignment in rapidly evolving pathogens and should be the subject of future research for diverse pathogens.

Materials and methods

The genomic data used in this study were submitted to a lineage assignment pipeline using five different versions of pangolin (Table 1), MAximum Parsimonious Likelihood Estimation (MAPLE), and Nextclade. The lineage assignments were then evaluated and compared through the methods described below (Fig. 1).

Data

We generated three datasets to compare the lineage assignment between pangoLEARN and pUShER as implemented in pangolin. The first (local dataset) consisted of 66,411 SARS-CoV-2 genomes collected in New York City (NYC) and the state of California (CA) with collection dates between the beginning of August 2021 and the end of November 2021: 15,862 genomes sampled in NYC by Department of Health and Mental Hygiene (DOHMH) Public Health Laboratory (PHL) and the NYC Pandemic Response Lab in addition to 50,549 genomes sampled in California via the California Department of Public Health (CDPH) COVIDNet sequencing effort. 15,393 NYC sequences and 45,326 CA sequences had a genome N percent content < 10. The second (global dataset 2021) was a random global dataset with 60,000 genomes from the National Center for Biotechnology Information (NCBI) with the same collection date range as the local dataset and sampled with equal amounts of genomes for each month with genome N percent content < 10. While random, this dataset was prone to bias due to differences in sequence deposition into NCBI impacted by factors such as total sequencing by country. The majority of genomes included came from just five countries: USA (31,633 sequences), England (19,097 sequences), Germany (4,677 sequences), Scotland (2,187 sequences), and Switzerland (1,662 sequences). The third (global dataset 2022) was a random global sample of 9,717 genomes from the Global Initiative on Sharing Avian Influenza Data (GISAID) collected in April 2022 with genome N percent content < 10. The composition of the third dataset skewed towards some of the same countries: USA (2,057 sequences), Denmark (1,730 sequences), England (1,280 sequences), and Germany (1,215 sequences). However, this sampling represents a more diverse subsample than previously attained (49 versus 32 countries) and more countries with more than 100 sequences included (14 versus 7 countries), potentially due to more countries contributing sequences. This third dataset was solely used for assessing pangolin v4 given that pangolin v3 was not trained to recognize the new lineages in this dataset.

Table 1. Pangolin and dependency versions. *With pangolin v4, pangoLEARN models became packaged into pangolin-data, which has a different versioning convention.

Pangolin	pangoLEARN	Scorpio	Constellations
v3.1.11	44,456	v0.3.12	v0.0.16
v3.1.14	44,467	v0.3.12	v0.0.16
v3.1.15	44,487	v0.3.13	v0.0.20
v3.1.16	44,509	v0.3.14	v0.0.21
v.4.0.2	1.2.133*	v0.3.16	v0.1.4

Lineage assignments

We performed lineage assignments to the dataset sequences using five versions of pangoLEARN and pUShER. Four versions spanned Pango designation v1.2.76–93, pangolin v.3.1.13, v.3.1.14, v.3.1.15, and v.3.1.16, where pangoLEARN uses a decision tree model and one version came from pangolin v.4.0.2/pangolin-data v1.2.133, where pangoLEARN uses a random forest model. For all the analyses, the option to assign lineages with designation hash was turned off with the option—skip-designation-hash to allow for a true comparison between both lineage designation methods.

Lineage assignment validation

In order to validate the lineage assignments from pangolin, we used an independent method with multiple sequence alignment followed by tree search. The multiple sequence alignment was performed using Multiple Alignment using Fast Fourier Transform (MAFFT) v7.486 (Katoh and Standley 2013) with options—anysymbol—keeplength—6merpair—addfragments on sequences from the local dataset that had less than 10 per cent unknown positions (*N*) as well as lineage consensus reference sequences (available at https://github.com/corneliusroemer/pango-seq

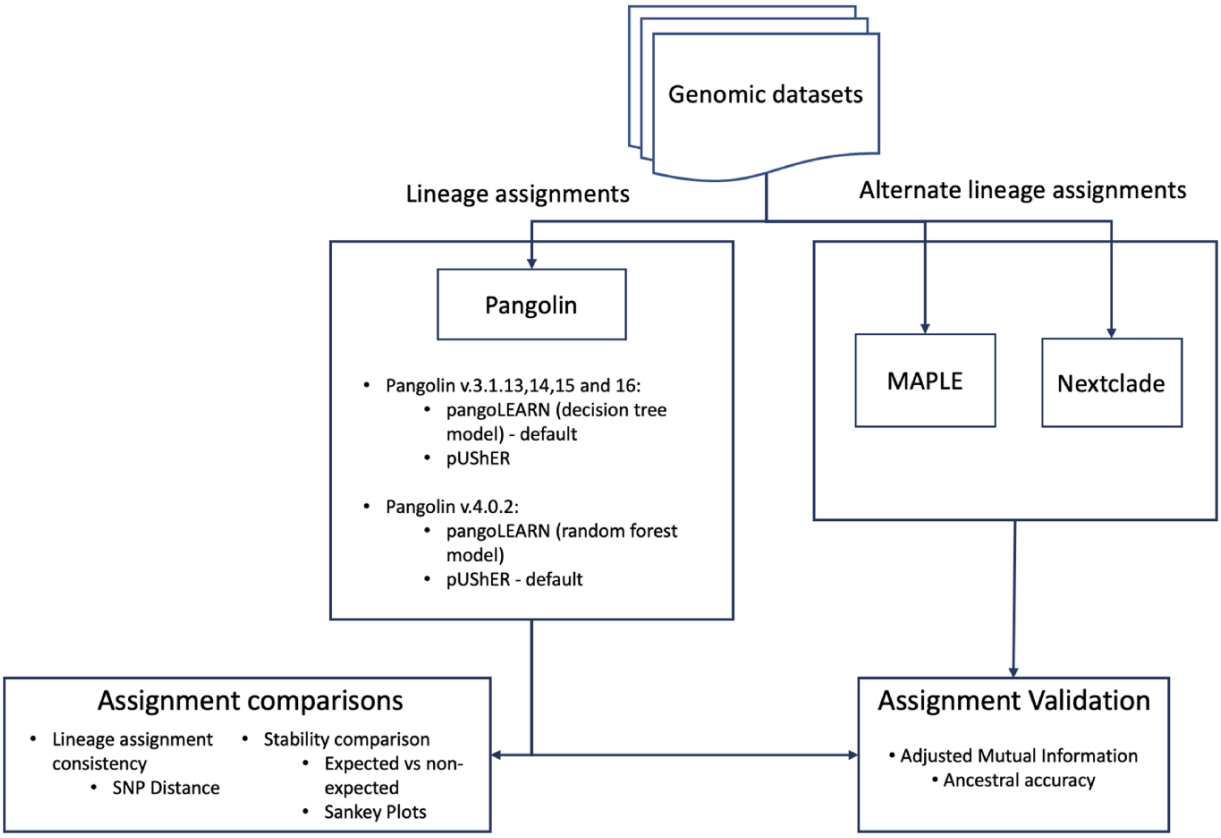

Figure 1. Lineage assignment and validation pipeline. Genomic datasets were run through pangolin and MAPLE for lineage assignments, assignment comparisons between pangolin assignments were performed for stability (expected versus non-expected) and lineage assignment consistency (SNP Distance), assignment validation. Assignment validation was performed comparing pangolin assignments with MAPLE assignments using Adjusted Mutual Information calculations and ancestral accuracy (lineage mis-assignments belonging to ancestral or descendent lineages or sublineages).

uences) for a total of 62,719 genome sequences. We implemented lineage assignment within the maximum likelihood phylogenetic software MAPLE (De Maio et al. 2023) version 0.3.4 (https://github.com/NicolaDM/MAPLE). This software was developed independently from pangolin and UShER. Lineage assignment in MAPLE is performed by inferring the joint phylogeny of reference and target genomes (using options—model UNREST—rateVariation—estimateSiteSpecificErrorRate), and then assigning each sample to the lineage whose reference genome is the closest direct ancestor. Sample A is interpreted as a direct ancestor of sample B if A has a phylogenetic distance (the sum of the branch lengths separating two nodes of the tree) of 0 from an internal node ancestral to B. During the MAPLE analysis, we chose to mask the untranslated ends of the sequences to the reference genome, specifically bases 1–265 and 29,674–29,903. This decision was made to ensure consistency with the input format used by pangolin for lineage assignment.

Disagreements between lineages assigned by MAPLE versus pangolin were scored by an in-house python script that determined whether or not one lineage was ancestral to the other, and computed the distance between the lineages as the number of edges separating the lineages on the MAPLE tree of all lineages. For example, B.1 was ancestral to B.1.2, with a distance of 1 (the edge from B.1 to B.1.2). B.1.3 and B.1.2 did not have an ancestral relationship; their common ancestor was B.1, and their distance was two (the edge from B.1 to B.1.3 and the edge from B.1 to B.1.2). We emphasize that this was not a genetic distance (i.e. the number of mutations that separate two lineages may not exactly correspond to this), but this comparison was appropriate for our analysis because we were evaluating correspondence within a lineage system. We used an in-house R script to perform an adjusted mutual information (AMI) comparison between the results from MAPLE and pangolin to see how consistently groups were recovered between the distinct methods.

In addition to MAPLE, we also validated the lineage assignments using Nextclade (Aksamentov et al. 2021). Nextclade CLI v1.11.0 was used with the SARS-CoV-2 dataset 2023–06-16T12:00:00Z. Disagreements between Nextclade and pangolin were scored by the same in-house python script described above.

Pangolin versions lineage assignment comparison

We created a list of expected versus non-permitted lineage assignments between each version based on the number of newly designated lineages. We then evaluated the relationship between genome coverage and number of lineage assignments across all versions of each method and number of non-permitted lineage changes (reassignment to a non-descendant lineage in subsequent pangolin versions). We also evaluated through Sankey diagrams of lineage assignment using the five different versions of pangolin (v3.1.13., v3.1.14, v3.1.15, v3.1.16, and v4.0.2) to look at the distinct pattern of assignment and reassignment (including non-permitted lineage changes) within pUShER and pangolin versions.

SNP distance between samples

A pairwise SNP distance matrix was created using snp-dists v.0.8.2 (https://github.com/tseemann/snp-dists). For each version of pangolin tested and each dataset, we calculated the SNP distances between all the sequences that were designated a lineage name.

Results and discussion

Dataset sampling

To evaluate the performance of pangolin calling by either pangoLEARN or pUShER, we built two initial datasets: a local US dataset that consists of 60,719 samples from CA and NYC and a 60,000 sample global dataset. The timeframe of sampling was

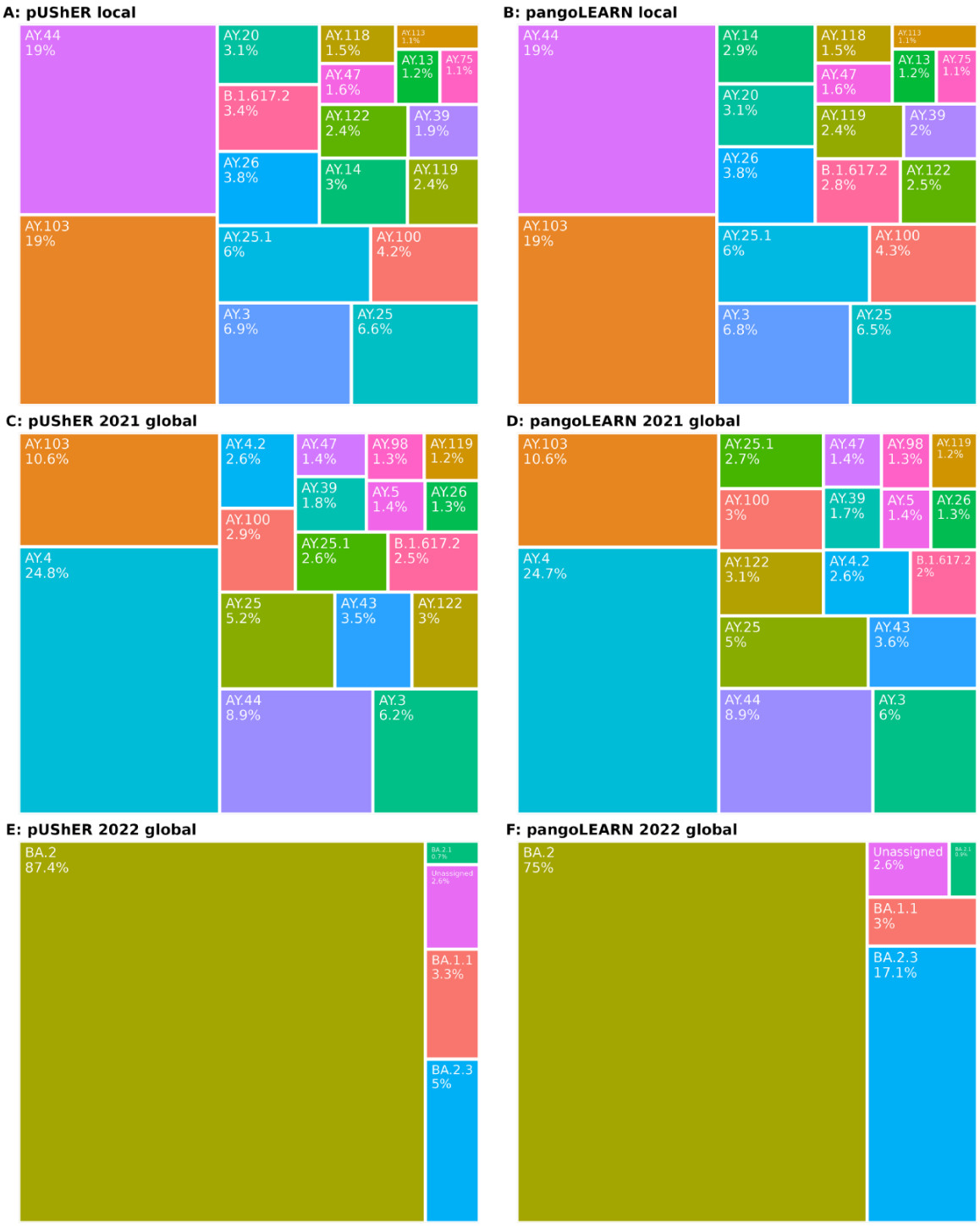

Figure 2. Lineage calls of datasets as determined by pUShER (A, C, E) and pangoLEARN (B, D, F) in Pangolin v.4.0.2. A/B: local dataset, C/D: 2021 global dataset, E/F: 2022 global dataset. For A–D, only lineages present at >1 per cent prevalence in the dataset are shown and for E–F, >0.01%.

August to November 2021 to avoid early sequencing quality concerns surrounding the newly designated Omicron strains. Therefore, samples included in this study were largely Delta sublineages (Fig. 2). The US dataset was more deeply sampled and could potentially magnify lineage assignment errors that apply only to a small set of samples, thus the performance against a global dataset was also necessary. It is worth noting that the majority of the global dataset consisted of sequences from the USA and England and was a biased subsampling of globally circulating strains.

There are differences in sublineage prevalence between the local and global datasets that reflect the effect of location on circulating variants. While countries can also enter different stages of the SARS-CoV-2 pandemic at different times, e.g. B.1.1.7 emerged in the UK before being found in other countries, the time period studied began in the middle of Delta's dominance (roughly May/June–December 2021) and was not a significant contributing factor.

After the release of pangolin v4, we created a third global dataset of 9,717 sequences from April 2022 to evaluate the performance of a newer methodology employed by pangoLEARN. This subset was still biased towards certain countries like the USA, England, Germany, and Denmark and reflected the waning of BA.1, which dominated during the early Omicron wave, and the subsequent takeover of BA.2. Additionally, it was a less diverse subset of sublineages than the previous datasets and performance on this subset may not necessarily have been generalizable.

Table 2. Agreement of calls of pangoLEARN and pUShER across the local and global datasets for different Pangolin versions (v3.1.13, v3.1.14, v3.1.16 and v4.0.2).

	Datasets		
Pangolin version	Local (%)	Global 2021 (%)	Global 2022 (%)
v3.1.13 (decision tree)	84.68	82.13	N/A
v3.1.14 (decision tree)	87.85	84.50	N/A
v3.1.15 (decision tree)	91.02	90.55	N/A
v3.1.16 (decision tree)	92.93	89.45	N/A
v4.0.2 (random forest)	97.28	97.35	86.90

Overall concordance and SNP distance

The concordance of the methods improved with each new model (Table 2). This was more apparent with the 2021 global dataset and was likely due to AY.4 prevalence. Overall, the two methods were highly concordant and differed mostly on sublineage calls.

Pangolin v4, released in April 2022, changed the pangoLEARN model to a random forest, fixed the previous calling errors of pangoLEARN (decision tree model), and was not significantly different from pUShER in any of the historical datasets. This was to be expected as similar sequences from those datasets were likely included in the training datasets for the new model. The pangoLEARN (random forest model) v4 had good concordance with pUShER on a contemporary dataset collected in April 2022, but it was much lower than with the historical datasets. The lower concordance could be largely attributed to pUShER assigning BA.2 and pangoLEARN BA.2.3. Assigning lineages with a newer version of pangoLEARN (random forest model) v.4.1.2 (UShER-v1.13 and pangoLEARN-v1.13) resolved the assignment in the favor of pUShER as pangoLEARN BA.2.3 calls decreased significantly from 1,662 to 392, while pUShER BA.2.3 calls changed minimally from 483 to 395. Thus, the new pangoLEARN (random forest model) method may still be prone to certain types of miscalling.

Because pangolin lineages are defined phylogenetically, we expected sequences that were given the same lineage designation to be more closely related than those of a different lineage.

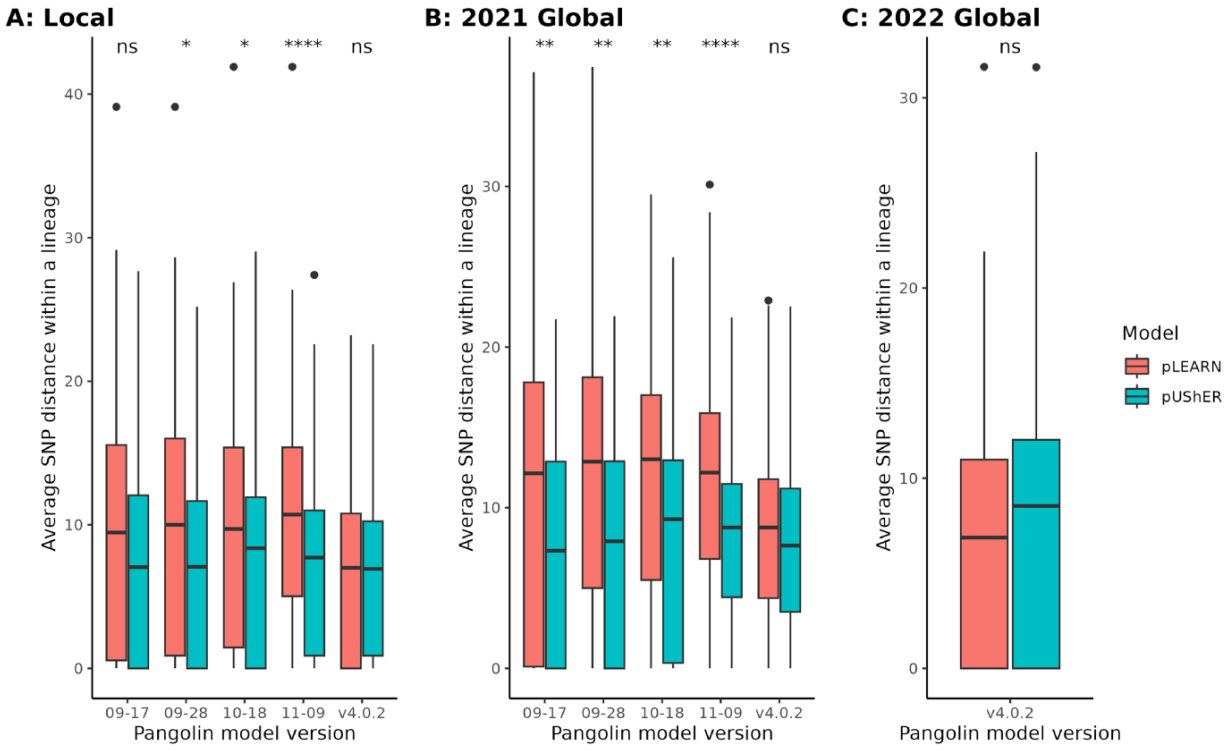

Figure 3. Average SNP distance between pangoLEARN and pUShER on Pangolin. (A) Local dataset, (B) 2021 global dataset, and (C) 2022 global dataset. ns = not statistically significant/* = 0.05/** = 0.01/**** = 0.0001. 09–17, 09–28, 10–18, and 11–09 used the pangoLEARN decision tree model, and v4.0.2 used the pangoLEARN random forest model.

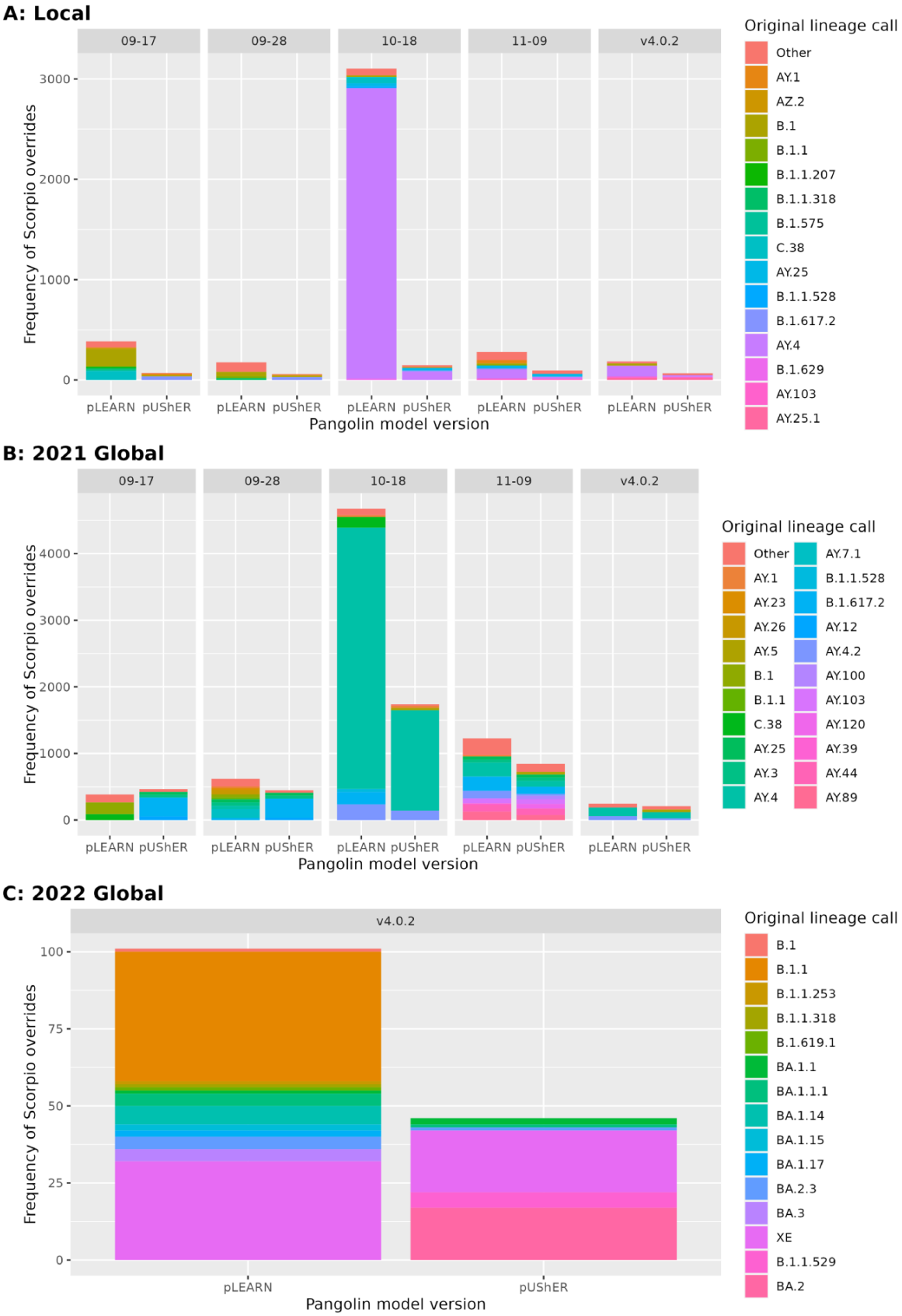

Figure 4. Number of Scorpio overrides between pangoLEARN and pUShER on Pangolin. (A) local dataset, (B) 2021 global dataset, (C) 2022 global dataset. 09–17, 09–28, 10–18, and 11–09 used the pangoLEARN decision tree model, and v4.0.2 used the pangoLEARN random forest model.

Pairwise SNP distances were calculated, and the average pairwise distance per lineage is shown in Fig. 3. It is evident that as the Delta wave progressed and more sublineages evolved, pUShER was able to call these newly defined sublineages with a lower average SNP distance between samples compared to pangoLEARN (decision tree model). This was true for both the local and global 2021 dataset with the difference between the two methods larger within the global. This was likely due to the high prevalence of AY.4 in the global 2021 dataset, which was known on a previous date to be overcalled by pangoLEARN (see https://github.com/cov-lineages/pango-designation/issues/221). In general, pangoLEARN (decision tree model) showed improvement over the course of the Delta wave by shrinking the average SNP distance between lineages with each successive model. Notably, pangoLEARN v4 (random forest model) did not differ significantly from pUShER in any of the datasets.

Scorpio analysis

Scorpio takes a set of lineage-defining 'constellations' with rules to classify each sequence by its specific mutations. It is manually curated and is limited to the WHO Variants of Concern, Variants of Interest, or Variants Under Monitoring (World Health Organization 2023). Prior to pangolin v4.1, when pangoLEARN or pUShER made an assignment that conflicted with Scorpio's assignment, pangolin overrode the pangoLEARN or pUShER assignment with the Scorpio assignment. This allowed pangolin to make higher accuracy assignments when new lineages emerged (pangoLEARN would initially lack sufficient training data for new lineages). However, this proved problematic with the emergence of BA.4 and BA.5 which saw Scorpio overriding correct assignments of these lineages and outputting BA.2 as the lineage (see https://github.com/cov-lineages/scorpio/issues/47). This Scorpio issue was due to early BA.4 and BA.5 having lower quality and lacking lineage defining mutations. However, due to the timeframe sampled, BA.4 and BA.5 sequences represent a small percentage of the data in our 2022 global dataset (<100 sequences). In addition, the latest versions of pangoLEARN and pUShER used in our study did not label these sequences as BA.4 and BA.5 at the time, thus, these sequences do not negatively affect our Scorpio analysis. No other large-scale problems with Scorpio have previously been documented. Therefore, Scorpio overrides in this study can be evaluated as erroneous calls made by pangoLEARN or pUShER.

In general, Scorpio was not a significant contributor to the overall accuracy of these methods as it was usually used in <1% of cases. Comparing the two methods, Scorpio overrode pangoLEARN calls more often than pUShER calls (Fig. 4). As expected, Scorpio was used to correct many incorrect AY.4 calls and accounted for the largest difference between the two methods (pangolin model 10–18). We observed pangoLEARN (decision tree model) erroneously calling B.1/B.1.1 sequences when pUShER did not have a similar problem. Given the timeframe of the specimens (Delta or Omicron wave), it was reasonable not to expect these sequences to be present, and this was corroborated by the genomic mismatches flagged by Scorpio in these sequences. B.1 miscalls were mostly present in pangolin model 9–17 and, to some extent, model 09–28. This reappeared as an issue in pangoLEARN (random forest model) v4 with B.1.1.

Initially, lineage categorization was performed using phylogenetic tree search methods such as IQTree. However, as the number of sequences grew, these methods became unfeasible. To address this issue, UShER became the go-to method for tree search and lineage categorization. This shift may explain why pUShER showed a higher degree of accuracy in lineage assignments compared to pangoLEARN.

Allowed reassignments

The Pango lineage system was explicitly designed to be updated with the SARS-CoV-2 pandemic as the virus continues to evolve (Rambaut et al. 2020). For that reason, some lineage reassignments were expected for a given genome sequence. For example, a sample originally designated in one lineage might be reassigned to a new daughter lineage of its original assignment. This occurs as an expected part of the Pango system when new lineages are designated. However, in some cases, a sample may be reassigned to a non-descendant lineage in subsequent versions of pangoLEARN or pUShER due to an error in the assignment approach. We will refer to such lineage reassignments as non-permitted lineage changes. Instability in lineage assignments might cause problems for interpretation that rely on precise and reliable lineage definitions for individual samples.

The consistency of assignments by pangoLEARN was inferior to pUShER. Even though 81 per cent of the sequences being assigned by pangoLEARN had a maximum of two calls across different pangolin versions, pUShER outperformed pangoLEARN by assigning 97 per cent of the sequences a maximum of two calls (Table 3). Although a large number of calls across different versions of pangoLEARN could be associated with the designation of new lineages, 27 per cent of the sequences analyzed with pangoLEARN presented at least one non-permitted change, while only 7 per cent of sequences assigned by pUShER present at least one

Table 3. Number of times (Khambaty, Bennett, and Shah 1994; Miranda et al. 1996; Thong, Puthucheary, and Pang 1998; Sandt et al. 2006; de Bernardi Schneider et al. 2019) that sequences were assigned a distinct lineage using pangoLEARN and pUShER across five different versions (v3.13, v3.v3.14, v3.15, v3.16, and v4).

Application\No. of calls	1	2	3	4	5
pangoLEARN	19,222 (29%)	34,463 (52%)	9537 (14%)	2515 (4%)	673 (1%)
pUShER	21,444 (32%)	43,229 (65%)	1619 (2%)	96 (>0%)	22 (>0%)

Table 4. Number of times (0-Thong, Puthucheary, and Pang 1998) that sequences were assigned non-permitted lineage changes using pangoLEARN and pUShER across five different versions (v3.13, v3.v3.14, v3.15, v3.16, and v4).

Application\Non-permitted changes	0	1	2	3	4
pangoLEARN	48,586 (73%)	10,706 (16%)	4170 (6%)	2072 (3%)	876 (1%)
pUShER	61,518 (93%)	3964 (6%)	670 (1%)	197 (>0%)	61 (>0%)

non-permitted change (Table 4 and Fig. 5). While pangoLEARN v4 (random forest model) was included in this analysis, the results were a reflection of the instability of pangoLEARN v3 (decision tree model) and cannot be generalized to v4.

Furthermore, the number of pangoLEARN lineage assignments for a given sequence seemed to be independent of the genome coverage, whereas pUShER assignments had a higher number of non-permitted lineage changes and consequently higher number of lineage assignments as the genome coverage decreased (Fig. 6). This reflects the expected higher phylogenetic placement uncertainty of less complete genomes.

Lineage assignment validation

As the pango lineage designation system is phylogenetic in nature, we wanted to benchmark the results of v4 pUShER and v4 pangoLEARN (random forest model) against a full maximum-likelihood phylogenetic method, MAPLE (De Maio et al. 2023) and Nextclade (Aksamentov et al. 2021). The MAPLE lineage assignment recovered 78.84% of the calls made by pUShER and 76.92% of the calls made by pangoLEARN. Nextclade recovered 97.52% of the calls made by pUShER and 96.09% of the calls by pangoLEARN (Table 5). The calculated AMI for all comparisons was above 0.85 indicating excellent recovery of similar

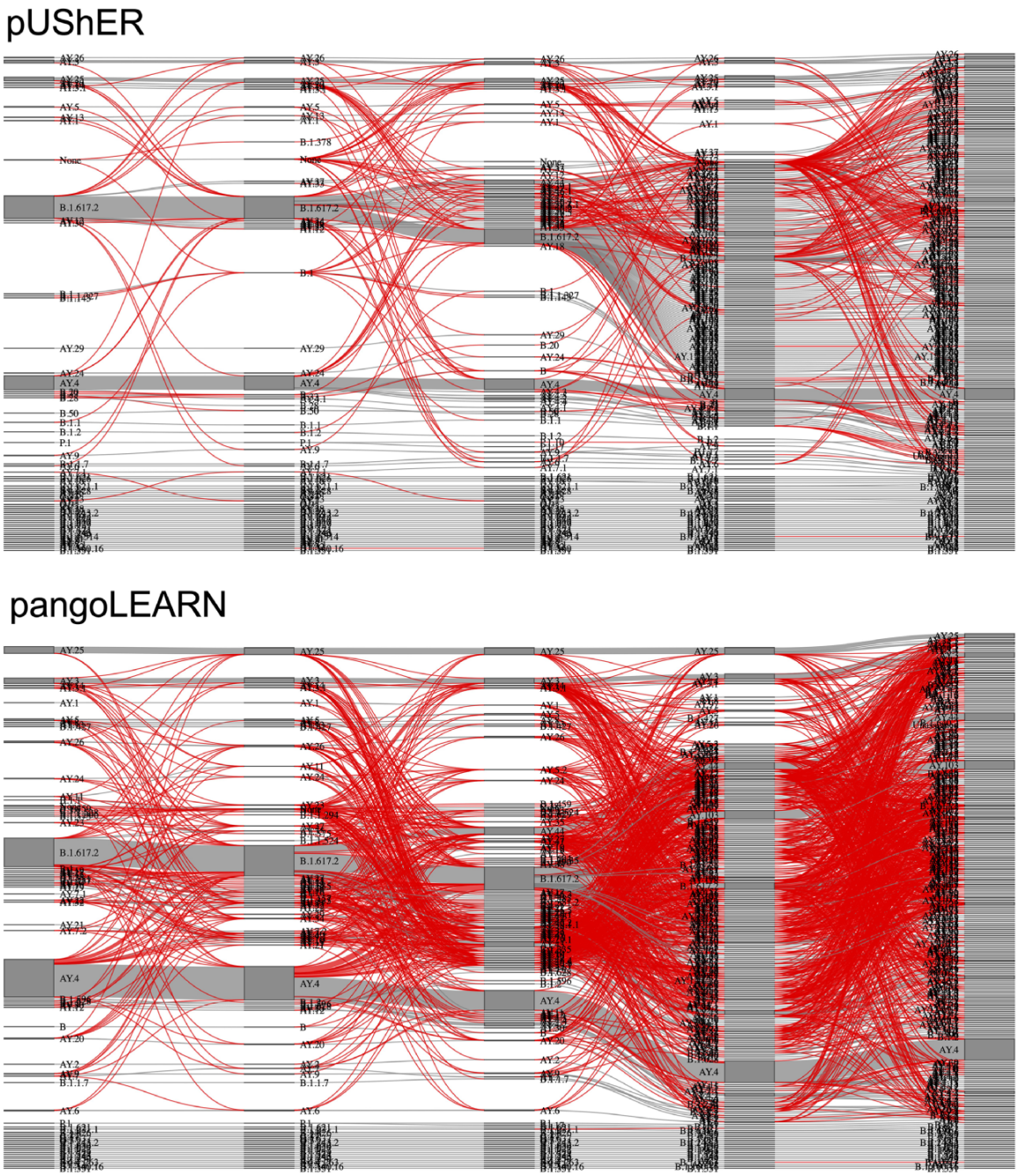

Figure 5. Sankey diagram of lineage assignment using pUShER (top) and pangoLEARN (bottom) across five different versions. Each column represents one version of pangolin in order of release (v3.1.13., v3.1.14, v3.1.15, v3.1.16, and v4.0.2). Red lines represent unexpected changes between versions. All sequences had N percent content < 10, and robustness estimates may differ for high ambiguous content. Two interactive html plots for this figure are available in the Supplementary Material, where the user can hover over each block within each column and identify the lineage and number of sequences that were labeled that lineage.

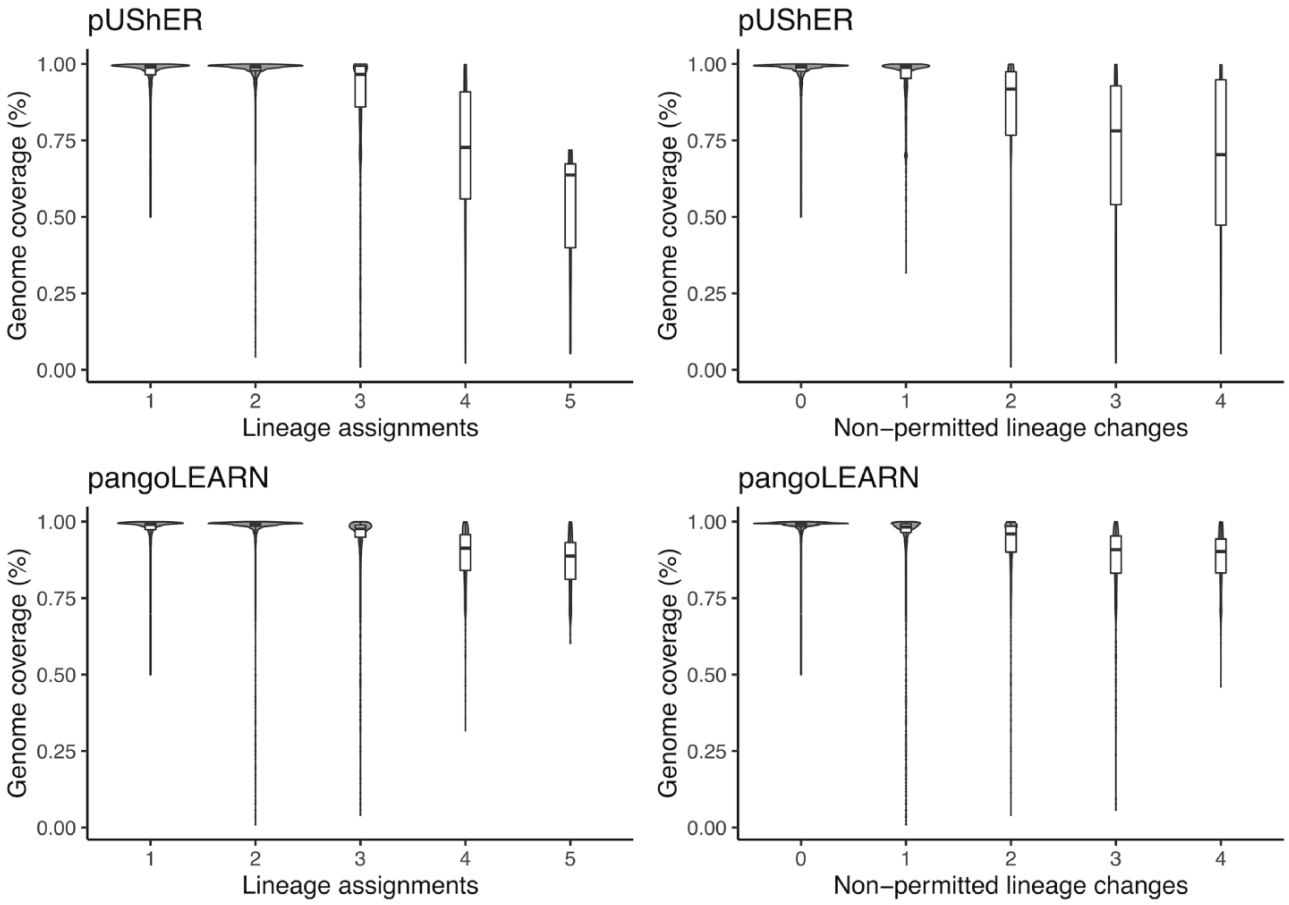

Figure 6. Violin plot of sample distribution based on genome coverage and pangoLEARN/pUShER lineage assignments for all CA and NYC samples. Top left: pUShER lineage assignment calls and reference genome coverage; Top right: pUShER lineage assignment non-permitted lineage changes and reference genome coverage; Bottom left: pangoLEARN lineage assignment calls and reference genome coverage; Bottom right: pangoLEARN lineage assignment non-permitted lineage changes and reference genome coverage.

Table 5. Validation of pangolin v.4 lineage assignments (pUShER and pangoLEARN) by comparison with MAPLE and Nextclade lineage assignments. Match and mismatch = Direct comparison of matches/mismatches of main call between the distinct methods. AMI (>0.90 = excellent recovery).

Method comparison	Matches	Mismatches	AMI
MAPLE vs pangoLEARN	46,707 (76.92%)	14,012 (23.08%)	0.88336081
MAPLE vs pUShER	47,872 (78.84%)	12,847 (21.16%)	0.906688939
Nextclade vs pangoLEARN	58,346 (96.09%)	2373 (3.91%)	0.948766048
Nextclade vs pUShER	59,216 (97.52%)	1503 (2.48%)	0.966070267

clusters of lineage assignments regardless of the specific lineage call made for each sequence, indicating consistency in the calls made by both validation methods in comparison to pUShER and pangoLEARN.

When looking into the mismatches, we found that 13,521 (96.6 per cent of mismatches) of MAPLE versus pangoLEARN (random forest model) mismatches were ancestrally related calls with distances of 1 or 2 sublineages between the calls, 13 (>0 per cent of mismatches) had a distance of 3+ sublineages, 470 (0.03 per cent) were siblings with distances of 2–4 to their common ancestor, and there was the presence of a mismatch due to a recombinant (XB) (Supplementary Table S1). For MAPLE versus pUShER, a similar ratio was found, with 12,510 (97.4%) being ancestrally related with distance of 1 or 2 sublineages between the calls, 4 (>0 per cent of mismatches) had a distance of 3+ sublineages, 324 (0.03 per cent) were siblings with distances of 2–4 to their common ancestor, and the presence of a mismatch due to a recombinant (XB) (Supplementary Table S2). Upon closer examination of the discrepancies between the pangolin lineage assignment methods and MAPLE, it becomes evident that the majority of mismatches occur between the AY.44/AY.26 and B.1.617.2 lineages. Specifically, 10,202 mismatches (79.41 per cent of total mismatches) are observed with the pUShER method, while 10,212 mismatches (72.88 per cent of total mismatches) are noted with the pangoLEARN method. Considering that the ancestral distance between these lineage assignments is just one sublineage, it leads us to hypothesize that MAPLE may have inaccurately positioned AY.44 and AY.26 as less ancestral lineages than they should have been.

For Nextclade versus pangoLEARN, a slighter different ratio was found, with 2,155 (90.9 per cent) being ancestrally related with the distance of 1 or 2 sublineages between the calls, 8 (0.3% of mismatches) had a distance of 3+ sublineages, 208 (8.8 per cent) were siblings with distances of 2–4 to their common ancestor, and the presence of two mismatches due to two recombinant (XB) (Supplementary Table S3). For Nextclade versus pUShER had a

similar result to the latter, with 1,367 (91 per cent) being ancestrally related with the distance of 1 or 2 sublineages between the calls, 4 (>0 per cent of mismatches) had a distance of 3+ sublineages and 132 (0.02 per cent) were siblings with distances of 2–4 to their common ancestor (Supplementary Table S4). The presence of over 90 per cent of the mismatches being ancestrally related with distances of 1 and 2 sublineages further the indication that there is a high concordance between all the methods analyzed.

Together, these results show that there is high concordance between pangolin methodology and two independent lineage assignment methods, giving confidence that pangolin can account for the phylogenetic structure underlying SARS-CoV-2 evolution. Nevertheless, the elevated number of mismatches between the pangolin methods and MAPLE suggests that further investigation is necessary to determine whether MAPLE would be inaccurately placing certain reference lineages during the tree search process.

Conclusions

Given the increased stability and reduced rate of non-permitted lineage reassignments by pUShER coupled with its higher reliability when analyzing high genome coverage/quality sequences compared to pangoLEARN v3 (decision tree model), we recommend that the pUShER option be selected as the first choice when assigning lineages to newly and previously sequenced genomes. While we hypothesize pangoLEARN v4 (random forest model) is more stable than v3, UShER has still been shown to have fewer Scorpio overrides and is likely to be as or more stable. We stress, however, that there are two important caveats to this recommendation.

First, the lineage system is explicitly phylogenetic, and recently, the Pango curation team has used a tree inferred with pUShER to assign new lineages. Thus, if there are systematic biases associated with phylogenetic inference in pUShER, then consistently inaccurate phylogenetic placements might create spurious lineage designations and assignments that appear to be consistent lineage calls. We consider this unlikely because the pUShER's accuracy for SARS-CoV-2 phylogenetic inference was quite high (Kramer et al. 2023; Turakhia et al. 2021).

Second, lineage assignments with pUShER had higher computational costs than with pangoLEARN; however, the compute costs associated with lineage assignment were relatively small compared to the total costs associated with producing a single-genome sequence. Furthermore, pUShER could efficiently exploit parallelism to decrease runtime. We believe that the advantages of increased stability of lineage assignments outweigh marginal additional computing costs except possibly when reanalyzing vast datasets on a regular basis as is done with large repositories such as GISAID. However, a complete reanalysis of 6 million genomes would cost approximately $43.44 on a typical cloud instance (see https://github.com/cov-lineages/pangoLEARN/issues/32#issuecomment-946937425) if efficiently exploiting multi-core architectures. This suggests that the cost is still a relatively minor consideration when choosing lineage assignment modes.

As of pangolin v4, pUShER is now the default option due to its accuracy and performance, which has been verified by benchmarking against the tree created by MAPLE and Nextclade. Changes have also been made to potentially decrease runtime through the implementation of an assignment cache (—add-assignment-cache and—use-assignment-cache). In addition, as of v4.1, Scorpio no longer overrides pUShER lineage assignments but continues to do so for pangoLEARN. Outbreak investigations are a case study of how pUShER's high accuracy and robustness can reduce superfluous resource consumption. Lineages called by pUShER can be trusted to cluster together on a phylogenetic tree and thus be genetically similar, avoiding chasing unrelated cases and maximizing resources allocated to contact tracing. Similarly, rapidly increasing sublineages can be scrutinized for higher fitness and/or immune evasion if the lineage calls can be trusted to be reliable. Thus, for newly emerging pathogens undergoing rapid evolution, our results suggest that phylogenetic placement is a superior option for lineage assignment than machine-learning methods.

Data availability

The dais manuscript are publicly available at NCBI and accessible upon request at GISAID. Sequence data with less than 90 per cent genome coverage that were not found in public databases were made available in a public GitHub repository. The list of the accessions, along with metadata and genomic sequences of the global and local datasets used in this study are available on GitHub at https://github.com/nychealth/COVID-consensus-genomes-pangolin-analysis.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We would like to acknowledge Rachel Colquhoun (University of Edinburgh) who has done most of the work on Scorpio/constellations as used by pangolin; the NYC DOHMH PHL staff; the CDPH California COVIDNet Team and these CDPH members: Dr Kathleen Jacobson, Dr Carol Glaser, Dr Mayuri Panditrao, Dr Christina Morales, Dr Nikki Baumrind, Elizabeth Baylis, Sabrina Gilliam; the University of California Office of the President, and COVIDNet WGS Lab Partners throughout California.

Funding

A.d.B.S. acknowledges support from the California Department of Public Health (Contract No. 20-11088). The findings and conclusions in this article are those of the authors and do not necessarily represent the views or opinions of the California Department of Public Health or the California Health and Human Services Agency. COVID sequencing in NYC was supported (in part) by the Epidemiology and Laboratory Capacity (ELC) for Infectious Diseases Cooperative Agreement (Grant Number: ELC DETECT (6NU50CK000517-01-07) funded by the Centers for Disease Control and Prevention (CDC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or the Department of Health and Human Services.

Conflict of interest: None declared.

References

Aksamentov, I. et al. (2021) 'Nextclade: Clade Assignment, Mutation Calling and Quality Control for Viral Genomes', *Journal of Open Source Software*, 6: 3773.

Cov-lineages. (2023) *Cov-lineages/scorpio: Serious Constellations of Reoccurring Phylogenetically-independent Origin* <https://github.com/cov-lineages/scorpio> accessed Sep 2023.

de Bernardi Schneider, A. et al. (2019) 'Updated Phylogeny of Chikungunya Virus Suggests Lineage-specific Rna Architecture', *Viruses*, 11: 798.

De Maio, N. et al. (2023) 'Maximum Likelihood Pandemic-scale Phylogenetics', *Nature Genetics.*, 55: 746–52.

Den Bakker, C. et al. (2014) 'Rapid Whole-genome Sequencing for Surveillance of Salmonella Enterica Serovar Enteritidis', *Emerging Infectious Diseases*, 20: 1306.

Dudas, G. et al. (2017) 'Virus Genomes Reveal Factors that Spread and Sustained the Ebola Epidemic', *Nature*, 544: 309–152017.

Durand, G. et al. (2018) 'Routine Whole-genome Sequencing for Outbreak Investigations of Staphylococcus Aureus in a National Reference Center', *Frontiers in Microbiology*, 9: 511.

Gilmour, M. W. et al. (2010) 'High-throughput Genome Sequencing of Two Listeria Monocytogenes Clinical Isolates during a Large Foodborne Outbreak', *BMC Genomics*, 11: 1–15.

Hadfield, J. et al. (2018) 'Nextstrain: Real-time Tracking of Pathogen Evolution', *Bioinformatics*, 34: 4121–3.

Jackson, B. R. et al. (2016) 'Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation', *Reviews of Infectious Diseases*, 63: 380–6.

Jajou, R. et al. (2018) 'A Predominant Variable-number Tandem-repeat Cluster of Mycobacterium Tuberculosis Isolates among Asylum Seekers in the Netherlands and Denmark, Deciphered by Whole-genome Sequencing', *Journal of Clinical Microbiology*, 56: e01100–17.

Jang, S. et al. (2005) 'PFGE-based Epidemiological Study of an Outbreak of Candida Tropicalis Candiduria: The Importance of Medical Waste as a Reservoir of Nosocomial Infection', *Japanese Journal of Infectious Diseases*, 58: 263.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Khambaty, F. M., Bennett, R. W., and Shah, D. B. (1994) 'Application of Pulsed-field Gel Electrophoresis to the Epidemiological Characterization of Staphylococcus Intermedius Implicated in a Food-related Outbreak', *Epidemiology & Infection*, 113: 75–81.

Kramer, A.M. et al. (2023) 'Online Phylogenetics with matOptimize Produces Equivalent Trees and is Dramatically More Efficient for Large SARS-CoV-2 Phylogenies than de novo and Maximum-Likelihood Implementations', *Systematic Biology*, 72: syad031.

Miranda, G. et al. (1996) 'Use of Pulsed-field Gel Electrophoresis Typing to Study an Outbreak of Infection Due to Serratia Marcescens in a Neonatal Intensive Care Unit', *Journal of Clinical Microbiology*, 34: 3138–41.

Moura, A. et al. (2017) 'Real-time Whole-genome Sequencing for Surveillance of Listeria Monocytogenes, France', *Emerging Infectious Diseases*, 23: 1462.

O'Toole, Á. et al. (2021) 'Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool', *Virus Evolution*, 7: veab064.

Rambaut, A. et al. (2020) 'A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 5: 1403–7.

——— et al. (2021) 'Addendum: A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology', *Nature Microbiology*, 6: 415–415.

Sandt, C. H. et al. The Key Role of Pulsed-field Gel Electrophoresis in Investigation of a Large Multiserotype and Multistate Food-borne Outbreak of Isalmonella/i Infections Centered in Pennsylvania. *Journal of Clinical Microbiology*, 44: 3208–12, September 2006. .

Scher, E., O'Toole, Á., and Rambaut, A. (2022) *Pangolearn Description*. <https://cov-lineages.org/resources/pangolin/pangolearn.html> accessed Feb 2023.

Stucki, D. et al. (2016) 'Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-incidence Country', *Journal of Clinical Microbiology*, 54: 1862–70.

Thong, K.-L., Puthucheary, S., and Pang, T. (1998) 'Outbreak of Salmonella enteritidis Gastroenteritis: Investigation by Pulsed-field Gel Electrophoresis', *International Journal of Infectious Diseases*, 2: 159–63.

Turakhia, Y. et al. (2021) 'Ultrafast Sample Placement on Existing Trees (Usher) Enables Real-time Phylogenetics for the SARS-Cov-2 Pandemic', *Nature Genetics.*, 53: 809–16.

World Health Organization. (2023) *Tracking SARS-Cov-2 variants*. <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>.

Zhang, T., Wu, Q., and Zhang, Z. (2020) 'Probable Pangolin Origin of SARS-Cov-2 Associated with the Covid-19 Outbreak', *Current Biology*, 30: 1346–51.