SARS-CoV-2 genomic diversity in households highlights the challenges of sequence-based transmission inference

Authors: Emily Bendall[1], Gabriela Paz-Bailey[2], Gilberto A. Santiago[2], Christina A. Porucznik[3], Joseph B. Stanford[3], Melissa S. Stockwell[4], Jazmin Duque[5], Zuha Jeddy[5], Vic Veguilla[2], Chelsea Major[2], Vanessa Rivera-Amill, PhD[6], Melissa A. Rolfes[2], Fatimah S. Dawood[2], Adam S. Lauring[1,7] *

Affiliations: [1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA; [2]Centers for Disease Control and Prevention; [3]Division of Public Health, Department of Family and Preventive Medicine, University of Utah School of Medicine; [4]Division of Child and Adolescent Health, Department of Pediatrics, Columbia University Vagelos College of Physicians and Surgeons, and Department of Population and Family Health, Mailman School of Public Health, Columbia University Irving Medical Center, New York, NY; [5]Abt Associates; [6]Ponce Health Sciences University/Ponce Research Institute; [7]Division of Infectious Diseases, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA;

*** Corresponding Author**: Adam S. Lauring, MS2 4742C, SPC 1621, 1137 Catherine Street, Ann Arbor, MI 48109; (734) 764-7731; alauring@med.umich.edu.

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Keywords: SARS-CoV-2; genomic epidemiology; transmission; household

Running Title: SARS-CoV-2 diversity in households

Abstract Word Count: 200

Main Text Word Count: 3009

Summary: High depth of coverage whole genome sequencing can identify SARS-CoV-2 transmission chains in settings where there is strong epidemiologic linkage but is not reliable as a stand-alone method for transmission inference.

ABSTRACT

Background: The reliability of sequence-based inference of SARS-CoV-2 transmission is not clear. Sequence data from infections among household members can define the expected genomic diversity of a virus along a defined transmission chain.

Methods: SARS-CoV-2 cases were identified prospectively among 2,369 participants in 706 households. Specimens with an RT-PCR cycle threshold ≤30 underwent whole genome sequencing. Intrahost single nucleotide variants (iSNV) were identified at ≥5% frequency. Phylogenetic trees were used to evaluate the relationship of household and community sequences.

Results: There were 178 SARS-CoV-2 cases in 706 households. Among 147 specimens sequenced, 106 yielded a whole genome consensus with coverage suitable for identifying iSNV. Twenty-six households had sequences from multiple cases within 14 days. Consensus sequences were indistinguishable among cases in 15 households, while 11 had ≥1 consensus that differed by 1-2 mutations. Sequences from households and the community were often interspersed on phylogenetic trees. Identification of iSNV improved inference in 2 of 15 households with indistinguishable consensus sequences and 6 of 11 with distinct ones.

Conclusions: In multiple infection households, whole genome consensus sequences differed by 0-1 mutations. Identification of shared iSNV occasionally resolved linkage, but the low genomic diversity of SARS-CoV-2 limits the utility of “sequence-only” transmission inference.

INTRODUCTION

RNA viruses evolve rapidly and accumulate mutations as outbreaks grow [1]. As a result, the evolutionary relationships among sequenced cases hold important information about the processes that drive epidemics [2]. For example, sequence data can help define transmission chains and outbreaks [3–5], the timing and location of viral introductions into communities [6–8], and larger patterns of spread [9–12]. Over the course of the COVID-19 pandemic, SARS-CoV-2 sequences have been used to infer transmission linkage in hospitals and other congregate settings [13–18]. Inferring these linkages with high confidence is necessary for subsequent studies of the biology of transmission and effectiveness of mitigation strategies.

To infer transmission, one can ask whether the sequences within a group of close contacts, such as a household, are more similar than sequences in the broader community. This approach depends on both the granularity of the sequence data and the amount of genomic diversity in the underlying community or meta-population. The relatedness of viral sequences identified from potential transmission chains versus community virologic surveillance is compared using phylogenetic trees of whole genome consensus sequences or clustering of transmission-associated sequences [2]. In the setting of insufficient community sampling and/or low genomic diversity, consensus trees can miss true linkages and identify false ones. Greater coverage sequencing can improve resolution by identifying intrahost single nucleotide variants (iSNV) in host-derived viral populations that have yet to achieve consensus levels, or >50% within-host frequency, along a transmission chain [19,20]. While these approaches have proven useful for influenza and other viruses, the reliability of sequence-based inference of SARS-CoV-2

46 transmission is less clear. For example, we and others have found that participants without
47 known epidemiologic linkage can share indistinguishable consensus sequences and even
48 minority (<50%) iSNV [21–23].
49
50 Households are ideal settings for studies of the biology and epidemiology of viral transmission.
51 Documentation of close contact and concurrent symptoms or test positivity provide strong
52 epidemiologic evidence of within-household transmission. Sequence data from infected
53 participants can therefore define the expected genomic diversity of a virus along a transmission
54 chain and inform sequence-based studies in other transmission settings, where epidemiologic
55 linkage may be uncertain. Here, we use whole genome sequencing of SARS-CoV-2 populations
56 from participants in two prospective household studies of COVID-19 that were conducted at
57 three sites. To assess the utility of SARS-CoV-2 sequence data as a tool for inferring
58 transmission, we used phylogenetic analysis of sequences from households with at least two
59 SARS-CoV-2 infection cases to assess the clustering of within-household sequences relative to
60 contemporaneous community sequences. We used iSNV to further resolve transmission
61 linkages in selected households.
62
63 **METHODS**
64 **Cohorts**
65 The Coronavirus Household Evaluation and Respiratory Testing (C-HEaRT) study enrolled
66 households in Utah (Salt Lake, Weber, Davis, Box Elder, Cache, Tooele, Wasatch, Summit, Utah,
67 and Iron Counties) and New York City [24] during August 2020 through February 2021 and

68 followed them with surveillance for SARS-CoV-2 infection during September 2020 through
69 August 2021. The Communities Organized for the Prevention of Arboviruses (COPA) was
70 expanded to include investigation of the epidemiology of COVID-19, creating the COCOVID
71 study, and recruited households in Ponce, Puerto Rico. For C-HEaRT, household eligibility
72 criteria included: ≥1 child aged 0-17 years, ≥75% of household members met individual level
73 eligibility (all members if a 2- or 3-person household), one adult member was willing to
74 complete monthly questionnaires, and adult members could communicate in English or
75 Spanish. Individual eligibility criteria included: anticipated residence in the household for ≥3
76 consecutive months, and willingness to complete study surveys, weekly symptom assessments,
77 and self-collect respiratory specimens. For COCOVID, household members were eligible if they
78 were aged ≥1 year, slept in the house ≥4 nights per week, had no definite plans to move in the
79 next year, and were willing and able to comply with study requirements. For both studies,
80 written informed consent (paper or electronic) was obtained from adults (aged >18 years in C-
81 HEaRT and >20 years in COCOVID). Parents or legal guardians of minor children provided
82 written informed consent on behalf of their children; older children (aged 12–17 years in C-
83 HEaRT and 7–20 years in COCOVID) also provided assent to study participation. The C-HEaRT
84 study protocol was reviewed and approved by the University of Utah Institutional Review Board
85 (IRB) as the single IRB for all collaborators. The COCOVID study protocol was reviewed and
86 approved by the Ponce Medical School Foundation, Inc. IRB.
87
88 **Sample Collection and Testing**

89 Participants were asked to self-collect (or parent/guardian-collect for children) mid-turbinate
90 nasal swabs every week, regardless of illness symptoms, and place the swabs in viral transport
91 media. Participants were also contacted by text message or email every week to ascertain if
92 they had COVID-19-like (CLI) or any other illness symptoms; they were asked to self-collect an
93 additional mid-turbinate flocked nasal swab once with onset of CLI symptoms. CLI was defined
94 as 1 or more of the following: fever or feverishness, cough, shortness of breath, sore throat,
95 diarrhea, muscle aches, chills, or change in taste or smell. Respiratory specimens were shipped
96 overnight to a central lab and tested using either the Quidel Lyra SARS-CoV-2 Assay or the
97 ThermoFisher Combo Kit platform. The assays were approved under Emergency Use
98 Authorization for the diagnosis of SARS-CoV-2 infection prior to use in this study. Test-positive
99 infections in the same household that were first detected by RT-PCR within 14 days of each
100 other (including those detected on the same date) were considered epidemiologically linked
101 and likely to have resulted from within-household transmission.
102
103 **SARS-CoV-2 Genomic Sequencing**
104 SARS-CoV-2 genomic sequencing was attempted on all specimens with an RT-PCR cycle
105 threshold (Ct) ≤30 on either the nucleocapsid protein 1 or 2 target. SARS-CoV-2 genomes were
106 sequenced as described previously [10]. Briefly, RNA was extracted from midturbinate nasal
107 swab specimens with the MagMax MVPII viral nucleic acid isolation kit on a Kingfisher Flex
108 (Thermofisher) and reverse transcribed with Lunascript (NEB). We amplified SARS-CoV-2 cDNA
109 in two pools using the ARTIC Network v3 primers and protocol. Amplicon pools were combined
110 in equal volumes for a given sample and purified with magnetic beads. Barcoded sequencing

111 libraries were prepared using the NEBNext ARTIC SARS-CoV-2 Library Prep Kit with magnetic
112 bead size selection. Individual barcoded sample libraries were pooled (up to 96) and sequenced
113 on an Illumina MiSeq (v2 chemistry, 2x250 cycles).
114
115 Reads were mapped to the Wuhan/Hu-1/2019 reference genome (GenBank MN908947.3) with
116 BWA-MEM [25]. We used iVar 1.2.1 [26] to trim ARTIC amplification primer sequences and to
117 determine consensus sequences using bases with >50% frequency and placing a designated
118 unknown base N at positions covered by fewer than 10 reads. Genomes with 29,000 or more
119 unambiguous bases (> 97% completeness) were used in downstream analysis. We identified
120 iSNV with iVar using the following parameters: sample with a minimum consensus genome
121 length of 29,000 bases; sample with an average genome sequencing coverage depth of greater
122 than 200 reads per position; iSNV frequency of 5–95%; read depth of 400 at iSNV sites with a
123 Phred score of >30; iVar p-value of <0.00001. We masked sites commonly affected by
124 sequencing errors in both consensus sequences and iSNV calls [27].
125
126 **Phylogenetic Analysis**
127 Consensus sequences for each household were placed on the global SARS-CoV-2 phylogenetic
128 tree using UShER [28]. The tree versions used were from the week of 14 February 2022 and
129 included over 7.8 million genome sequences from GISAID, GenBank, COG-UK and CNCB. The
130 level of genomic sampling of the state or territory of each study site (Figure 1) was estimated
131 with subsampler [10] using case data and GISAID submission data. Subtrees were initially
132 constructed with 30 samples and then reconstructed with additional samples as needed to

133 visualize all genomes from a household in a single subtree (e.g., when samples existed within
134 large clusters of indistinguishable samples). The JSON files for each master tree and subtree are
135 available in Supplemental Dataset 1 and can be visualized in the auspice viewer at
136 https://auspice.us/. Trees were annotated and edited in FigTree using the subtree.nwk files
137 generated by UShER.
138
139 **Data and materials availability**
140 The consensus genomes that we generated for this study are publicly available on
141 https://github.com/lauringlab/SARS-CoV-2_Household_Diversity Accessions. Those for the
142 community sequences (largely from GISAID) can be found in the pdf tree files
143 ("<Household_ID>.pdf") in Supplemental Dataset 1. Laboratories responsible for submissions
144 are acknowledged in Supplemental Table 1. Analysis code for the generation of consensus
145 sequences and phylogenetic analysis are available at https://github.com/lauringlab/SARS-CoV-
146 2_Household_Diversity.
147
148 **RESULTS**
149 The C-HEaRT (Utah and New York City) and COCOVID (Puerto Rico) studies performed active
150 surveillance for SARS-CoV-2 infection and CLI in 706 households with 2,369 participants (Table
151 1). During September 2020 through August 2021, the cumulative incidence of SARS-CoV-2
152 infection was 11% [96/842 participants in 41/190 (22%) households under surveillance] at the
153 Utah site and 7% [33/499 participants in 13/135 (10%) households] at the New York City site;

154 during June 2020 through September 2021, cumulative incidence was 5% at the Puerto Rico site
155 (49/1028 infections detected in 28/381 households).
156
157 Of the 191 participants with SARS-CoV-2 infections in these households, 147 (77%) in 70
158 households had samples with a Ct value <30 that were processed for whole genome
159 sequencing, of whom, 106 (72%) had samples that were successfully sequenced to sufficient
160 breadth and depth of coverage (see Methods). Of the 706 households at the three sites, 56
161 included ≥2 participants who were test-positive within a 14-day period, suggestive of within-
162 household transmission (Table 1). Twenty-six households had high quality sequence data on ≥2
163 of these contemporaneous infections. The SARS-CoV-2 clades and lineages identified (Table 2)
164 were the same among participants of the same household and reflect viruses circulating in the
165 corresponding time periods in the United States (www.outbreak.info)(Table 2).
166
167 We first used phylogenetic analysis of whole genome sequences to infer transmission linkage
168 within these 26 households. We used Usher [28] to obtain local sequences for each household
169 and to place household sequences simultaneously on a phylogenetic tree. Over 7.8 million
170 whole genome SARS-CoV-2 sequences were available at the time of this analysis. Because most
171 of these contextual sequences were from GISAID, we estimated the level of sampling at each
172 study site over time by dividing the number of GISAID sequences by the number of reported
173 cases [10]. Sampling of locally circulating viruses was low in 2020 (<2% cases sequenced) and
174 increased at all three sites beginning in early 2021 (>5% cases sequenced, Figure 1). In 2022,
175 Utah was generally better sampled (~10-30%) than New York or Puerto Rico (5-15%).

In 15 out of the 26 households that we studied, the consensus sequences of all cases were indistinguishable and grouped together on their respective trees (representative trees in Figure 2, additional trees in Supplemental Figures 1-4). Given the epidemiologic linkage in the same household, these can be considered sequence-confirmed transmission events. However, we also found several trees in which these monophyletic groupings also included indistinguishable, contemporaneous sequences from non-household members within the community of the same locality (see trees in Figure 2). In two of the New York households, there were many such sequences during a B.1.526 (Iota) variant wave (Supplemental Figures 1 and 2). Therefore, even with a modest level of sampling (Figure 1), it is not uncommon to find indistinguishable viral sequences from participants at the same region and time who presumably lack a documented epidemiologic linkage.

In 11 households, the consensus sequences of the virus from one or more household members differed at 1-2 positions over the nearly 30kb genome. This is not uncommon in transmission chains, particularly ones that are longer or in which the samples are collected 7-14 days apart (see Table 2 for time span). In nearly all cases, the trees from these households demonstrated linkage and/or an ancestor/descendant relationship for the viral sequences (representative trees in Figure 3, additional trees in Supplemental Figure 5). In some cases, the household lineages were phylogenetically distinct from contemporaneous local sequences (e.g., UT2, UT4). These tree structures supported transmission linkage, but low sampling of community cases makes it hard to rule out missed linkages between members of the household and the larger

198 community. Indeed, there were some households in which there were sequences from the
199 larger community included among the same branches of the within-household sequences (e.g.,
200 UT10, PR3). As above, the low genomic diversity of SARS-CoV-2 and modest sampling of
201 community cases made it difficult to define a threshold to effectively rule in or rule out
202 transmission.
203
204 We next determined whether transmission inference could be improved by identifying iSNV
205 that were shared among members of a household. These would manifest as polymorphic sites
206 where the alternative allele, or mutation, is present but not fixed in the transmission chain.
207 While there was just one household where two participants shared a minority iSNV (PR5, Figure
208 4), several had iSNV in at least one individual at a consensus level (i.e., frequency >0.5), but that
209 had not yet achieved fixation (i.e., frequency >0.95). In the 15 households with
210 indistinguishable consensus sequences, each of the two participants in households NY5 and PR5
211 shared an iSNV that was consensus level, but not fixed. In three (UT12, UT5, UT3) out of the 11
212 households with distinct consensus sequences (Table 2), the consensus differences were due to
213 one or more participants having a non-reference iSNV that achieved consensus level, but not
214 fixation. Household UT4 had two participants with consensus level iSNV (Figure 4). In household
215 PR3, there was one site where one out of four members had a consensus level iSNV and
216 another member had this as a fixed mutation.
217
218 **DISCUSSION**

219 We evaluated the utility of SARS-CoV-2 sequence data in transmission inference using data
220 from two studies of household cohorts at three sites. In the household setting, where at least
221 two incident infections occurring within 14 days of one another are strongly suggestive of
222 transmission, we found that sequencing generally confirmed transmission linkage. The whole
223 genome consensus sequences of participants within a household were nearly always
224 indistinguishable or differed by one mutation. In some cases, these links were further
225 supported by the identification of iSNV shared among members of the household. Out of the 26
226 households evaluated, there was just one (UT10) in which the high average number of
227 consensus differences (two) and absence of shared iSNV called linkage into doubt. Importantly,
228 we frequently found multiple sequences from the community that were indistinguishable to
229 those within the household with even modest sampling (<5%) over the course of the pandemic.
230 This highlights the limits of "sequence-only" inference of transmission in hospitals or other
231 congregate settings where epidemiologic linkage is less certain.
232
233 Strengths of the study include our reliance on samples from active surveillance of longitudinal
234 cohorts and our use of quality controlled, deep sequencing. With weekly sampling of all
235 participants from a household, we were able to identify asymptomatic or mildly symptomatic
236 cases and avoid some of the bias of case-ascertainment studies, in which cases are recruited
237 based on a test-positive index. Together with our use of contemporaneous community
238 specimens collected from participants not in the households but from the same site, our data
239 provide a valuable benchmark for the expected SARS-CoV-2 diversity in households relative to
240 that in the community. The cohorts are also drawn from diverse geographic areas with varied

241 household sizes and composition [21,29]. Our assessment of viral diversity is strengthened by
242 our criteria for identifying consensus and minority iSNV [22]. The low observed diversity in this
243 study, in part, reflects the stringent thresholds applied to the sequence data . This conservative
244 approach reduces sequencing errors, which can be systematic and lead to incorrect
245 ascertainment of shared iSNV among unrelated participants [21–23,30].
246
247 This study had several notable limitations. First, we were relatively stringent in our criteria for
248 identifying iSNV and therefore may have underascertained shared diversity in the rare (<5%)
249 variant fraction within households. Second, given the limited number of households with
250 sequenced cases, we were unable to formulate a statistically robust approach to sequence-
251 based inference with clear cut-offs and associated positive and negative predictive values. Case-
252 ascertained cohorts or contact tracing studies offer a more efficient way to capture and
253 sequence many putative transmission pairs and will be useful as a setting in which to further
254 develop this approach. Third, while we believe that our data provide an important framework
255 for interpreting sequence data in studies of SARS-CoV-2 transmission, data from households
256 may not translate completely to hospitals and other congregate living settings, which may differ
257 in case density, contact frequency, and force of infection. Fourth, we assumed that household
258 cases testing positive for SARS-CoV-2 within 14 days of one another were linked by
259 transmission. If these cases represented distinct introductions into the household, we could
260 overestimate expected within-household diversity. Fifth, it is possible, but in our opinion
261 unlikely, that some of the community cases in our analysis actually had an epidemiologic
262 linkage to participants in these households.

263

264 Despite the limitations identified in this study, integration of sequence and epidemiologic data
265 can be a powerful approach to studies of SARS-CoV-2 transmission. In settings where there is
266 strong epidemiologic linkage among cases (e.g., known exposure or clear temporal and spatial
267 association), indistinguishable consensus sequences with or without shared iSNV should be
268 confirmatory. In these situations, single mutation differences among consensus sequences in a
269 cluster are not uncommon; mutations can fix along a transmission chain, particularly longer
270 ones over a greater timespan. However, if epidemiologic linkage is less certain, sequence
271 identity can only confirm transmission if the metapopulation is highly sampled and genetically
272 diverse. For example, early in the pandemic when circulating SARS-CoV-2 diversity was low,
273 many inpatients and employees in hospitals were found to share indistinguishable consensus
274 sequences, and even iSNV, without any apparent epidemiologic linkage [22,31]. This contrasts
275 with other studies of hospital outbreaks where the combination of contact tracing and
276 sequence data confirmed suspected transmission chains and identified new ones. We expect
277 that future studies of transmission in households, hospitals, and other congregate settings will
278 benefit from Bayesian methods, which can integrate epidemiologic and sequence data for
279 improved inference [32].

280

281 **ACKNOWLEDGMENTS**

282 We thank the participants in the C-HEaRT and COCOVID cohorts and all GISAID submitting
283 laboratories. We acknowledge Anderson Britto for developing and suggesting the subsampler
284 tool used to generate Figure 1.

285 **REFERENCES**

286 1. Lauring AS. Within-Host Viral Diversity: A Window into Viral Evolution. Annu Rev Virol
287 **2020**; 7:63–81.

288 2. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing
289 and big data are transforming epidemiology. Trends Microbiol **2014**; 22:282–291.

290 3. Lemieux JE, Siddle KJ, Shaw BM, et al. Phylogenetic analysis of SARS-CoV-2 in Boston
291 highlights the impact of superspreading events. Science **2021**; 371:eabe3261.

292 4. Siddle KJ, Krasilnikova LA, Moreno GK, et al. Transmission from vaccinated individuals in a
293 large SARS-CoV-2 Delta variant outbreak. Cell **2022**; 185:485-492.e10.

294 5. Zeller M, Gangavarapu K, Anderson C, et al. Emergence of an early SARS-CoV-2 epidemic in
295 the United States. Cell **2021**; 184:4939-4952.e15.

296 6. Worobey M, Pekar J, Larsen BB, et al. The emergence of SARS-CoV-2 in Europe and North
297 America. Science **2020**; 370:564–570.

298 7. du Plessis L, McCrone JT, Zarebski AE, et al. Establishment and lineage dynamics of the
299 SARS-CoV-2 epidemic in the UK. Science **2021**; 371:708–712.

300 8. Candido DS, Claro IM, de Jesus JG, et al. Evolution and epidemic spread of SARS-CoV-2 in
301 Brazil. Science **2020**; 369:1255–1260.

302 9. Viana R, Moyo S, Amoako DG, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron
303 variant in southern Africa. Nature **2022**; 603:679–686.

304 10. Alpert T, Brito AF, Lasek-Nesselquist E, et al. Early introductions and transmission of SARS-
305 CoV-2 variant B.1.1.7 in the United States. Cell **2021**; 184:2595-2604.e13.

306 11. Valesano AL, Fitzsimmons WJ, Blair CN, et al. SARS-CoV-2 Genomic Surveillance Reveals
307 Little Spread From a Large University Campus to the Surrounding Community. Open Forum
308 Infect Dis **2021**; 8:ofab518.

309 12. Aggarwal D, Warne B, Jahun AS, et al. Genomic epidemiology of SARS-CoV-2 in a UK
310 university identifies dynamics of transmission. Nat Commun **2022**; 13:751.

311 13. Lucey M, Macori G, Mullane N, et al. Whole-genome Sequencing to Track Severe Acute
312 Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Transmission in Nosocomial Outbreaks.
313 Clin Infect Dis **2021**; 72:e727–e735.

314 14. Francis RV, Billam H, Clarke M, et al. The Impact of Real-Time Whole-Genome Sequencing
315 in Controlling Healthcare-Associated SARS-CoV-2 Outbreaks. J Infect Dis **2022**; 225:10–18.

316 15. Meredith LW, Hamilton WL, Warne B, et al. Rapid implementation of SARS-CoV-2
317 sequencing to investigate cases of health-care associated COVID-19: a prospective
318 genomic surveillance study. Lancet Infect Dis **2020**; 20:1263–1271.

319 16. Hamilton WL, Fieldman T, Jahun A, et al. Applying prospective genomic surveillance to
320 support investigation of hospital-onset COVID-19. Lancet Infect Dis **2021**; 21:916–917.

321 17. Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 Infections and
322 Transmission in a Skilled Nursing Facility. N Engl J Med **2020**; 382:2081–2090.

323 18. Addetia A, Crawford KHD, Dingens A, et al. Neutralizing Antibodies Correlate with
324 Protection from SARS-CoV-2 in Humans during a Fishery Vessel Outbreak with a High
325 Attack Rate. J Clin Microbiol **2020**; 58:e02107-20.

326 19. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic processes
327 constrain the within and between host evolution of influenza virus. eLife **2018**; 7:e35962.

328 20. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission
329 Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol **2017**; 186:1209–1216.

330 21. Braun KM, Moreno GK, Wagner C, et al. Acute SARS-CoV-2 infections harbor limited
331 within-host diversity and transmit via tight transmission bottlenecks. PLOS Pathog **2021**;
332 17:e1009849.

333 22. Valesano AL, Rumfelt KE, Dimcheff DE, et al. Temporal dynamics of SARS-CoV-2 mutation
334 accumulation within and across infected hosts. PLOS Pathog **2021**; 17:e1009499.

335 23. Tonkin-Hill G, Martincorena I, Amato R, et al. Patterns of within-host genetic diversity in
336 SARS-CoV-2. eLife **2021**; 10:e66857.

337 24. Dawood FS, Porucznik CA, Veguilla V, et al. Incidence Rates, Household Infection Risk, and
338 Clinical Characteristics of SARS-CoV-2 Infection Among Children and Adults in Utah and
339 New York City, New York. JAMA Pediatr **2022**; 176:59.

340 25. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
341 ArXiv13033997 Q-Bio **2013**; Available at: http://arxiv.org/abs/1303.3997. Accessed 31 July
342 2020.

343 26. Grubaugh ND, Gangavarapu K, Quick J, et al. An amplicon-based sequencing framework
344 for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol
345 **2019**; 20:8.

346 27. De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. Masking strategies
347 for SARS-CoV-2 alignments. Virological. 2020; Available at:
348 https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480. Accessed 10
349 March 2022.

350 28. Turakhia Y, Thornlow B, Hinrichs AS, et al. Ultrafast Sample placement on Existing tRees
351 (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet **2021**;
352 53:809–816.

353 29. Walter KS, Kim E, Verma R, et al. Shared within-host SARS-CoV-2 variation in households.
354 medRxiv **2022**; :2022.05.26.22275279.

355 30. Lythgoe KA, Hall M, Ferretti L, et al. SARS-CoV-2 within-host diversity and transmission.
356 Science **2021**; :eabg0821.

357 31. Braun KM, Moreno GK, Buys A, et al. Viral Sequencing to Investigate Sources of SARS-CoV-
358 2 Infection in US Healthcare Personnel. Clin Infect Dis **2021**; 73:e1329–e1336.

359 32. Lindsey BB, Villabona-Arenas ChJ, Campbell F, et al. Characterising within-hospital SARS-
360 CoV-2 transmission events using epidemiological and viral genomic data across two
361 pandemic waves. Nat Commun **2022**; 13:671.

362

363

364 **Table 1:** Distribution of SARS-CoV-2 test-positive cases across cohorts and households

365

	New York City	Utah	Puerto Rico
Number of Households	135	190	381
Number of Participants	499	842	1028
Median household size (range)	4 (2-9)	4 (2-10)	2 (1-6)
Number of unique SARS-CoV-2-positive cases[a]	33	96	49
Households with 1 case	3	17	3
Households with 2 cases[b]	5	7	5
Households with 3 cases[b]	3	5	8
Households with 4 cases[b]	0	4	6
Households with 5 cases[b]	1	4	5
Households with 6 cases[b]	1	0	1
Households with 7 cases[b]	0	1	0
# of cases with sequence data / # sequenced	28/29	52/86	26/32

[a] total number over study period
[b] only households with cases testing positive within 14 days of each other

366

367 **Table 2:** Households with two or more incident SARS-CoV-2 infections within a 14- day period
368

Household	Number of Specimens Sequenced	Date First Specimen	Date Last Specimen	Nextclade Clade[a]	PANGO lineage[b]	Mean Consensus Diff[c]
PR1	3	9/2/20	9/8/20	20C	B.1.426	0
UT1	3	10/14/20	10/15/20	20G	B.1.2	0
PR2	4	10/24/20	10/30/20	20C	B.1.588	0.5
UT2	4	11/17/20	11/24/20	20B	B.1.1	0.5
UT3	6	11/24/20	12/2/20	20G	B.1.2	1.4
UT4	4	11/30/20	12/7/20	20B	B.1.1	1.5
UT5	2	12/2/20	12/3/20	20A	B.1.400	1
UT6	3	12/3/20	12/10/20	20G	B.1.2	0.67
PR3	4	12/3/20	12/11/20	20B	B.1.1.486	1.17
UT7	3	12/15/20	12/28/20	20A	B.1.596	0
UT8	2	12/28/20	1/1/21	21C (Epsilon)	B.1.429	0
UT9	2	1/12/21	1/14/21	20A	B.1.400	0
NY1	3	1/29/21	2/4/21	21F (iota)	B.1.526	0.67
PR4	2	2/8/21	2/8/21	20A	B.1.240	0
NY2	3	2/9/21	2/9/21	21F (Iota)	B.1.526	0
NY3	3	2/11/21	2/18/21	20C	B.1.582	0
NY4	2	2/21/21	3/5/21	21F (iota)	B.1.526	0
NY5	2	2/23/21	2/23/21	21F (Iota)	B.1.526	0
NY6	6	2/24/21	3/11/21	20C	B.1.637	1.13
UT10	2	3/1/21	3/15/21	21C (Epsilon)	B.1.427	2
NY7	3	3/3/21	3/16/21	20C	B.1.637	0
NY8	2	3/22/21	4/4/21	21F (Iota)	B.1.526	0
PR5	2	3/23/21	3/30/21	20B	R.1	0
PR6	3	4/23/21	4/23/21	20I (Alpha,V1)	Q.4	0
UT11	3	7/26/21	8/10/21	21J (Delta)	AY.44	0
UT12	2	8/4/21	8/4/21	21J (Delta)	AY.44	1

a Defined using nextclade, https://clades.nextstrain.org
b Defined using pango, https://cov-lineages.org/resources/pangolin.html

c Total number of pairwise unambiguous consensus differences between sequences divided by total number of sequences in a household

369
370

371 **FIGURE LEGENDS**

372

373 **Fig. 1**: **Cases and sampling density.** Columns show the number of SARS-CoV-2 infections (left y-
374 axis) in households from New York (NY, top), Puerto Rico (PR, middle), and Utah (UT, bottom)
375 cohorts by epiweek (x-axis). The sampling density (line) for community genomes in each state
376 or territory (right y-axis) was estimated as the proportion of cases with sequences available on
377 GISAID.

378

379 **Fig. 2: Phylogenetic trees of sequences from households where all participants had**
380 **indistinguishable consensus sequences.** Shown are four representative trees. Trees from 11
381 other households are shown in Supplemental Figures 1-4. Each tree is labeled with the
382 household identifier (NY = New York, UT = Utah, PR = Puerto Rico). The tips of household
383 sequences are colored cyan and those from non-household participants in the same community
384 in the same state or territory (2 letter abbreviation) are colored magenta. All other tips are
385 colored black. The collection date for each specimen is indicated. Genetic distance is
386 represented by the bar and corresponds to one mutation.

387

388 **Fig. 3: Phylogenetic trees of sequences from households where participants had distinct**
389 **consensus sequences.** Shown are four representative trees. Trees from 7 other households are
390 shown in Supplemental Figure 5. Each tree is labeled with the household identifier (NY = New
391 York, UT = Utah, PR = Puerto Rico). The tips of household sequences are colored cyan and those
392 from the same state or territory (2 letter abbreviation) are colored magenta. All other tips are

393 colored black. The collection date for each sample is indicated. Genetic distance is represented
394 by the bar and corresponds to one mutation.
395
396 **Fig. 4: Shared single nucleotide polymorphisms within households.** Each panel shows one of
397 the 8 households in which members shared a polymorphic site. The frequency (5-95%) of the
398 indicated mutation (relative to the Wuhan/Hu-1 reference) is shown on the y-axis and the
399 individual/sequence identifier is shown on the x-axis. Shared variants that did and did not lead
400 to a consensus level difference between household members are shown as circles and
401 diamonds, respectively. Mutations that are fixed (>95% frequency) are shown as squares.
402
403 **Supplemental Figures 1-4: Phylogenetic trees of sequences from households where all**
404 **participants had indistinguishable consensus sequences.** Each tree is labeled with the
405 household identifier (NY = New York, UT = Utah, PR = Puerto Rico). The tips of household
406 sequences are colored cyan and those from the same state or territory (2 letter abbreviation)
407 are colored magenta. All other tips are colored black. The collection date for each sample is
408 indicated. Genetic distance is represented by the bar and corresponds to one mutation.
409
410 **Supplemental Figure 5: Phylogenetic trees of sequences from seven households where**
411 **participants had distinct consensus sequences.** Each tree is labeled with the household
412 identifier (NY = New York, UT = Utah, PR = Puerto Rico). The tips of household sequences are
413 colored cyan and those from the same state or territory (2 letter abbreviation) are colored

414 magenta. All other tips are colored black. The collection date for each sample is indicated.

415 Genetic distance is represented by the bar and corresponds to one mutation.

416

417 **Supplemental Table 1:** Submitting laboratories for GISAID sequences used in this study. GISAID

418 identifiers can be found in the unedited trees in the Supplemental Dataset.

419

Figure 1

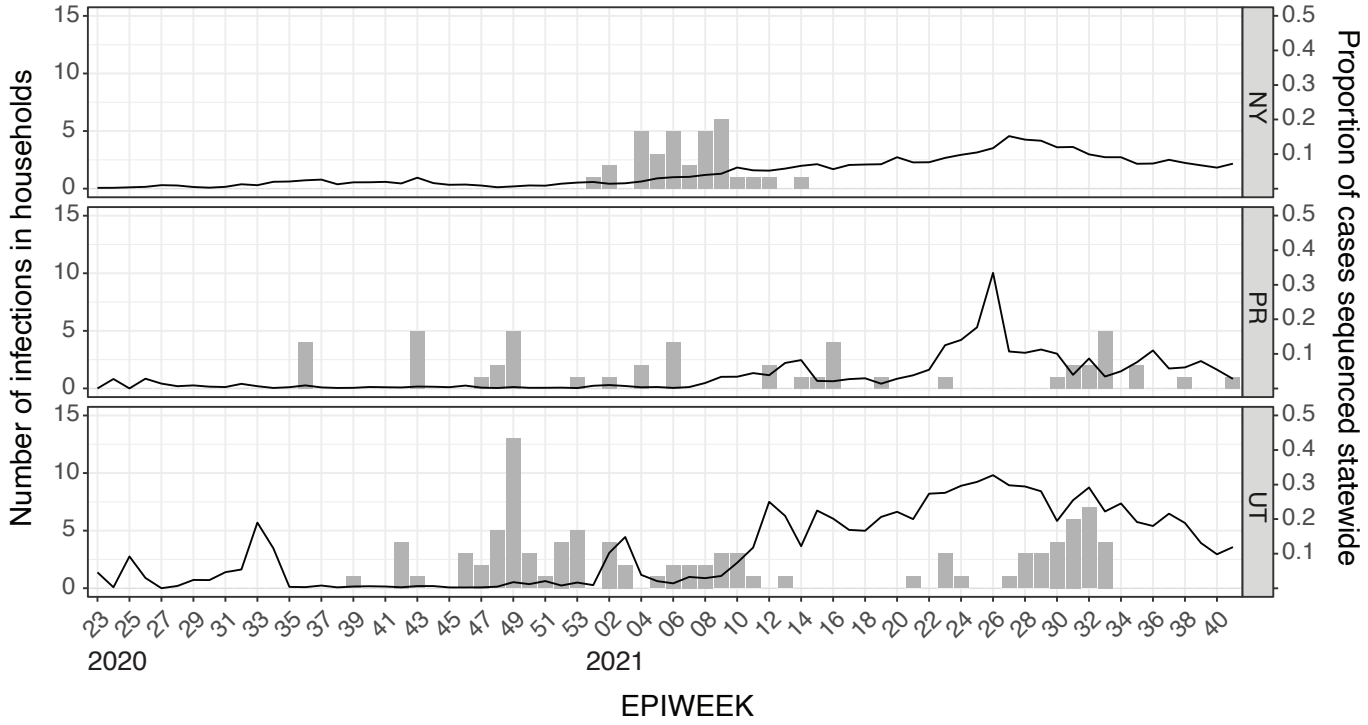

Number of infections in households
Proportion of cases sequenced statewide
NY
PR
UT
0
5
10
15
0.1
0.2
0.3
0.4
0.5
23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 02 04 06 08 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
2020
2021
EPIWEEK

Figure 2

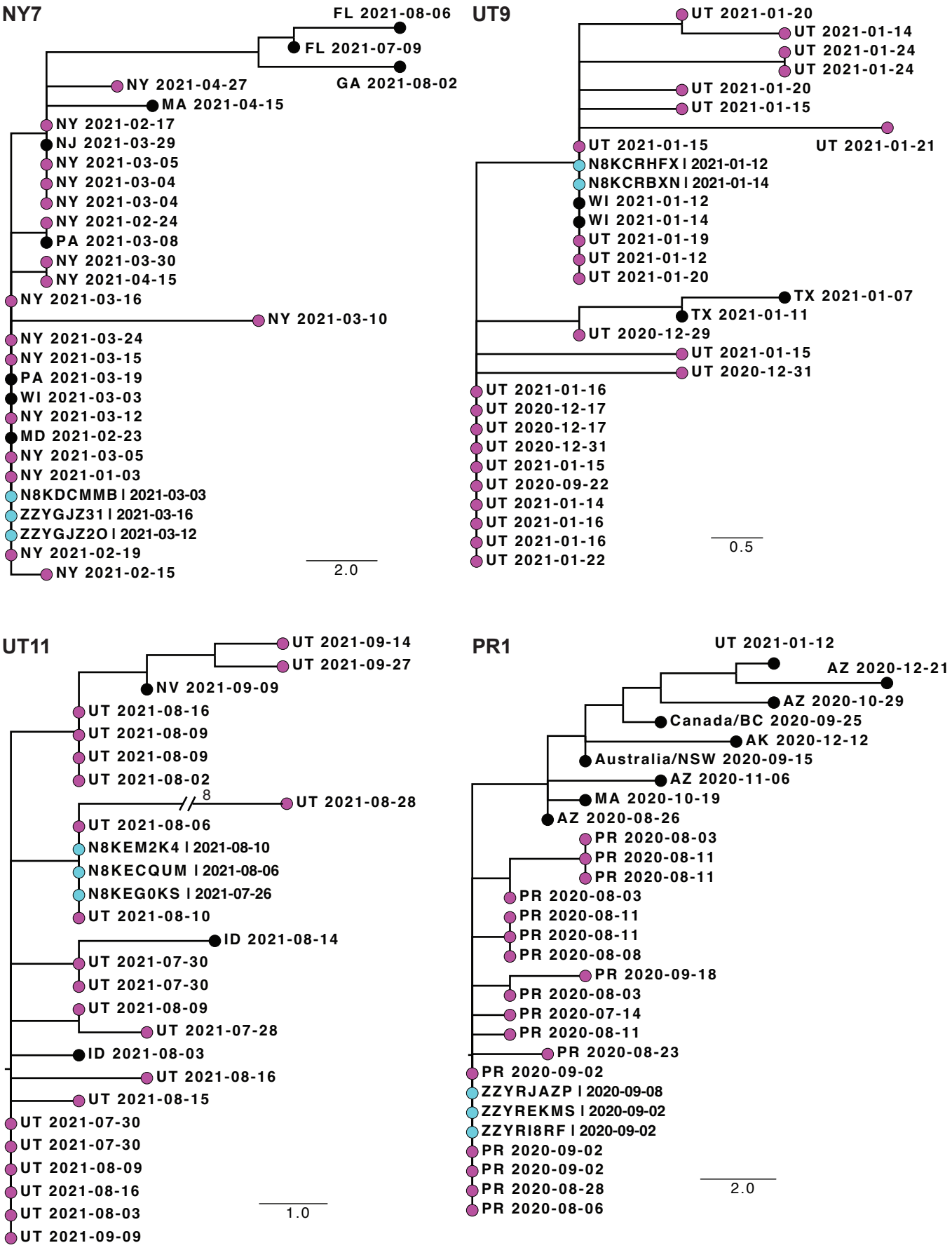

NY7
FL 2021-08-06
FL 2021-07-09
GA 2021-08-02
NY 2021-04-27
MA 2021-04-15
NY 2021-02-17
NJ 2021-03-29
NY 2021-03-05
NY 2021-03-04
NY 2021-03-04
NY 2021-02-24
PA 2021-03-08
NY 2021-03-30
NY 2021-04-15
NY 2021-03-16
NY 2021-03-10
NY 2021-03-24
NY 2021-03-15
PA 2021-03-19
WI 2021-03-03
NY 2021-03-12
MD 2021-02-23
NY 2021-03-05
NY 2021-01-03
N8KDCMMB | 2021-03-03
ZZYGJZ31 | 2021-03-16
ZZYGJZ2O | 2021-03-12
NY 2021-02-19
NY 2021-02-15
2.0
UT9
UT 2021-01-20
UT 2021-01-14
UT 2021-01-24
UT 2021-01-24
UT 2021-01-20
UT 2021-01-15
UT 2021-01-21
UT 2021-01-15
N8KCRHFX | 2021-01-12
N8KCRBXN | 2021-01-14
WI 2021-01-12
WI 2021-01-14
UT 2021-01-19
UT 2021-01-12
UT 2021-01-20
TX 2021-01-07
TX 2021-01-11
UT 2020-12-29
UT 2021-01-15
UT 2020-12-31
UT 2021-01-16
UT 2020-12-17
UT 2020-12-17
UT 2020-12-31
UT 2021-01-15
UT 2020-09-22
UT 2021-01-14
UT 2021-01-16
UT 2021-01-16
UT 2021-01-22
0.5
UT11
UT 2021-09-14
UT 2021-09-27
NV 2021-09-09
UT 2021-08-16
UT 2021-08-09
UT 2021-08-09
UT 2021-08-02
8
UT 2021-08-28
UT 2021-08-06
N8KEM2K4 | 2021-08-10
N8KECQUM | 2021-08-06
N8KEG0KS | 2021-07-26
UT 2021-08-10
ID 2021-08-14
UT 2021-07-30
UT 2021-07-30
UT 2021-08-09
UT 2021-07-28
ID 2021-08-03
UT 2021-08-16
UT 2021-08-15
UT 2021-07-30
UT 2021-07-30
UT 2021-08-09
UT 2021-08-16
UT 2021-08-03
UT 2021-09-09
1.0
PR1
UT 2021-01-12
AZ 2020-12-21
AZ 2020-10-29
Canada/BC 2020-09-25
AK 2020-12-12
Australia/NSW 2020-09-15
AZ 2020-11-06
MA 2020-10-19
AZ 2020-08-26
PR 2020-08-03
PR 2020-08-11
PR 2020-08-11
PR 2020-08-03
PR 2020-08-11
PR 2020-08-11
PR 2020-08-08
PR 2020-09-18
PR 2020-08-03
PR 2020-07-14
PR 2020-08-11
PR 2020-08-23
PR 2020-09-02
ZZYRJAZP | 2020-09-08
ZZYREKMS | 2020-09-02
ZZYRI8RF | 2020-09-02
PR 2020-09-02
PR 2020-09-02
PR 2020-08-28
PR 2020-08-06
2.0

Figure 3

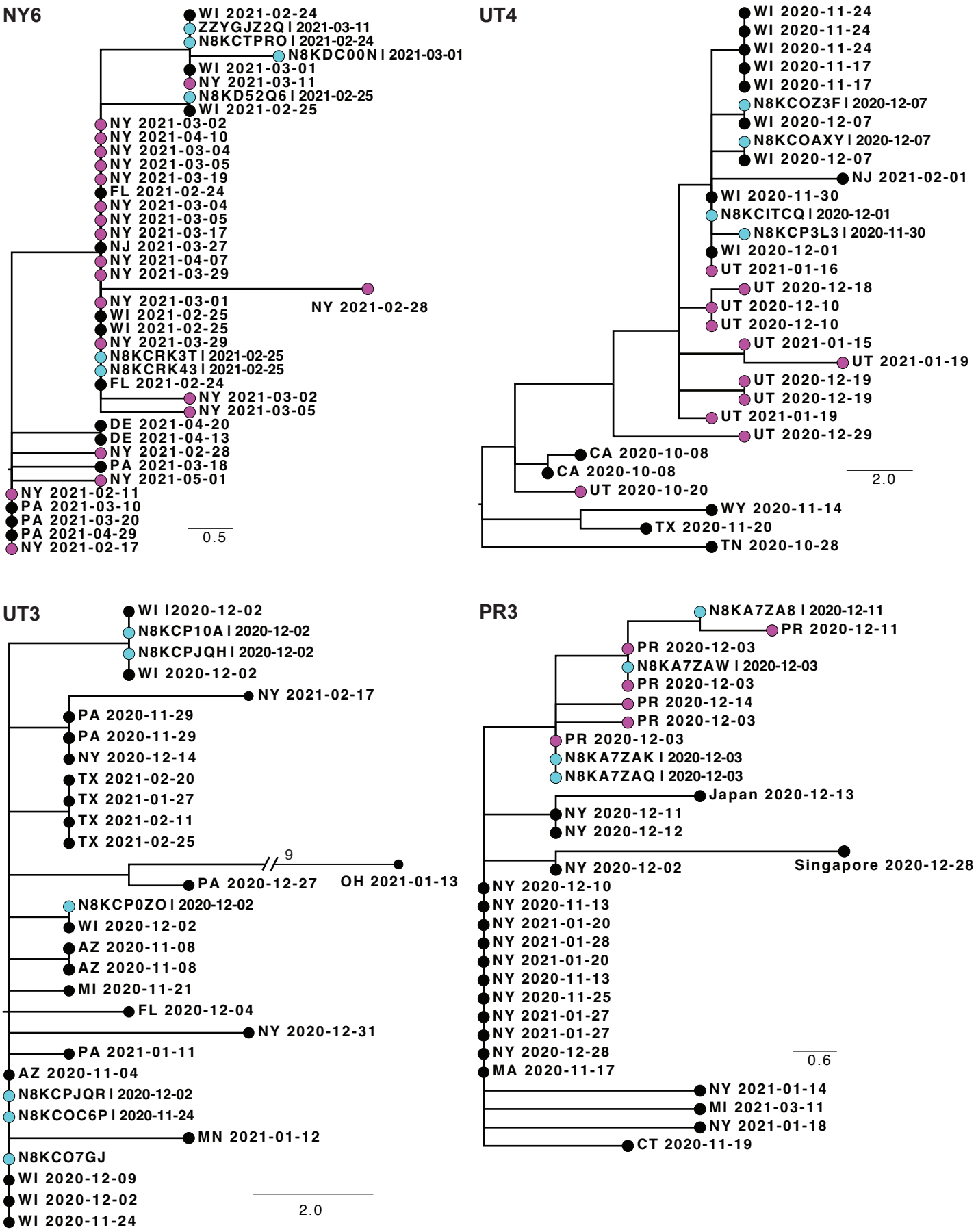
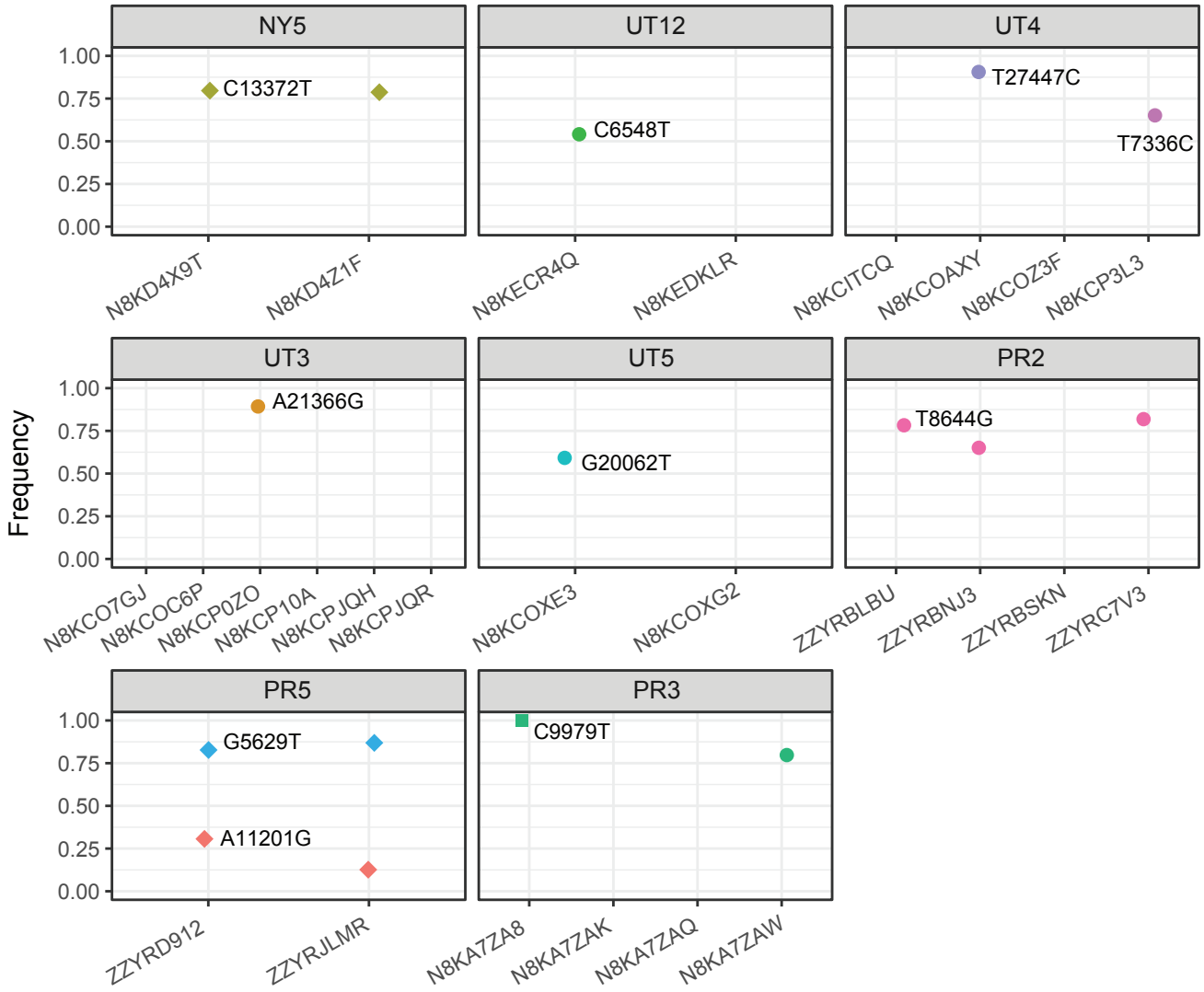
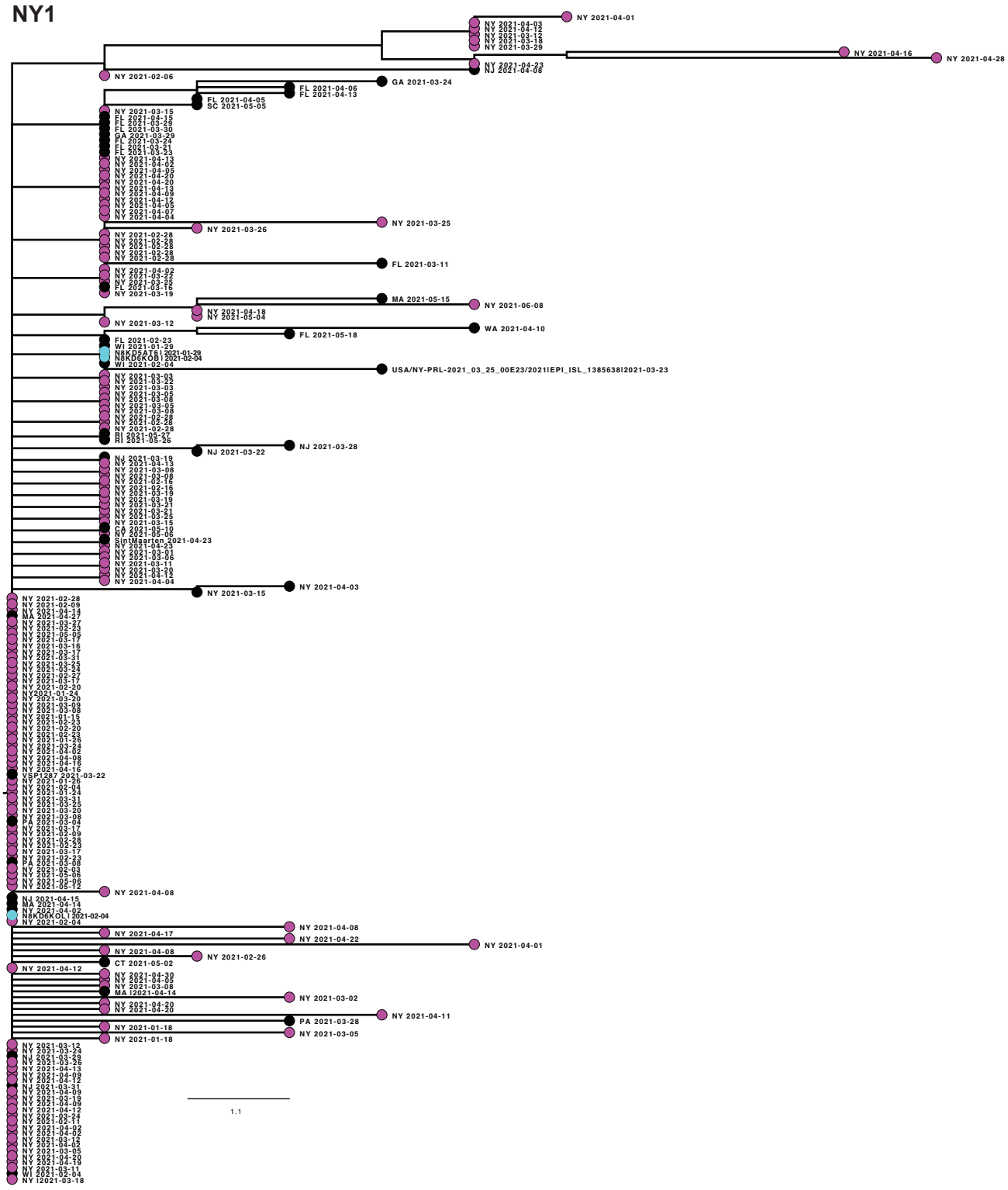

NY6
WI 2021-02-24
ZZYGJZ2Q | 2021-03-11
N8KCTPRO | 2021-02-24
N8KDC00N | 2021-03-01
WI 2021-03-01
NY 2021-03-11
N8KD52Q6 | 2021-02-25
WI 2021-02-25
NY 2021-03-02
NY 2021-04-10
NY 2021-03-04
NY 2021-03-05
NY 2021-03-19
FL 2021-02-24
NY 2021-03-04
NY 2021-03-05
NY 2021-03-17
NJ 2021-03-27
NY 2021-04-07
NY 2021-03-29
NY 2021-02-28
NY 2021-03-01
WI 2021-02-25
WI 2021-02-25
NY 2021-03-29
N8KCRK3T | 2021-02-25
N8KCRK43 | 2021-02-25
FL 2021-02-24
NY 2021-03-02
NY 2021-03-05
DE 2021-04-20
DE 2021-04-13
NY 2021-02-28
PA 2021-03-18
NY 2021-05-01
NY 2021-02-11
PA 2021-03-10
PA 2021-03-20
PA 2021-04-29
NY 2021-02-17
0.5
UT4
WI 2020-11-24
WI 2020-11-24
WI 2020-11-24
WI 2020-11-17
WI 2020-11-17
N8KCOZ3F | 2020-12-07
WI 2020-12-07
N8KCOAXY | 2020-12-07
WI 2020-12-07
NJ 2021-02-01
WI 2020-11-30
N8KCITCQ | 2020-12-01
N8KCP3L3 | 2020-11-30
WI 2020-12-01
UT 2021-01-16
UT 2020-12-18
UT 2020-12-10
UT 2020-12-10
UT 2021-01-15
UT 2021-01-19
UT 2020-12-19
UT 2020-12-19
UT 2021-01-19
UT 2020-12-29
CA 2020-10-08
CA 2020-10-08
UT 2020-10-20
WY 2020-11-14
TX 2020-11-20
TN 2020-10-28
2.0
UT3
WI |2020-12-02
N8KCP10A | 2020-12-02
N8KCPJQH | 2020-12-02
WI 2020-12-02
NY 2021-02-17
PA 2020-11-29
PA 2020-11-29
NY 2020-12-14
TX 2021-02-20
TX 2021-01-27
TX 2021-02-11
TX 2021-02-25
9
OH 2021-01-13
PA 2020-12-27
N8KCP0ZO | 2020-12-02
WI 2020-12-02
AZ 2020-11-08
AZ 2020-11-08
MI 2020-11-21
FL 2020-12-04
NY 2020-12-31
PA 2021-01-11
AZ 2020-11-04
N8KCPJQR | 2020-12-02
N8KCOC6P | 2020-11-24
MN 2021-01-12
N8KCO7GJ
WI 2020-12-09
WI 2020-12-02
WI 2020-11-24
2.0
PR3
N8KA7ZA8 | 2020-12-11
PR 2020-12-11
PR 2020-12-03
N8KA7ZAW | 2020-12-03
PR 2020-12-03
PR 2020-12-14
PR 2020-12-03
PR 2020-12-03
N8KA7ZAK | 2020-12-03
N8KA7ZAQ | 2020-12-03
Japan 2020-12-13
NY 2020-12-11
NY 2020-12-12
Singapore 2020-12-28
NY 2020-12-02
NY 2020-12-10
NY 2020-11-13
NY 2021-01-20
NY 2021-01-28
NY 2021-01-20
NY 2020-11-13
NY 2020-11-25
NY 2021-01-27
NY 2021-01-27
NY 2020-12-28
MA 2020-11-17
0.6
NY 2021-01-14
MI 2021-03-11
NY 2021-01-18
CT 2020-11-19

Figure 4

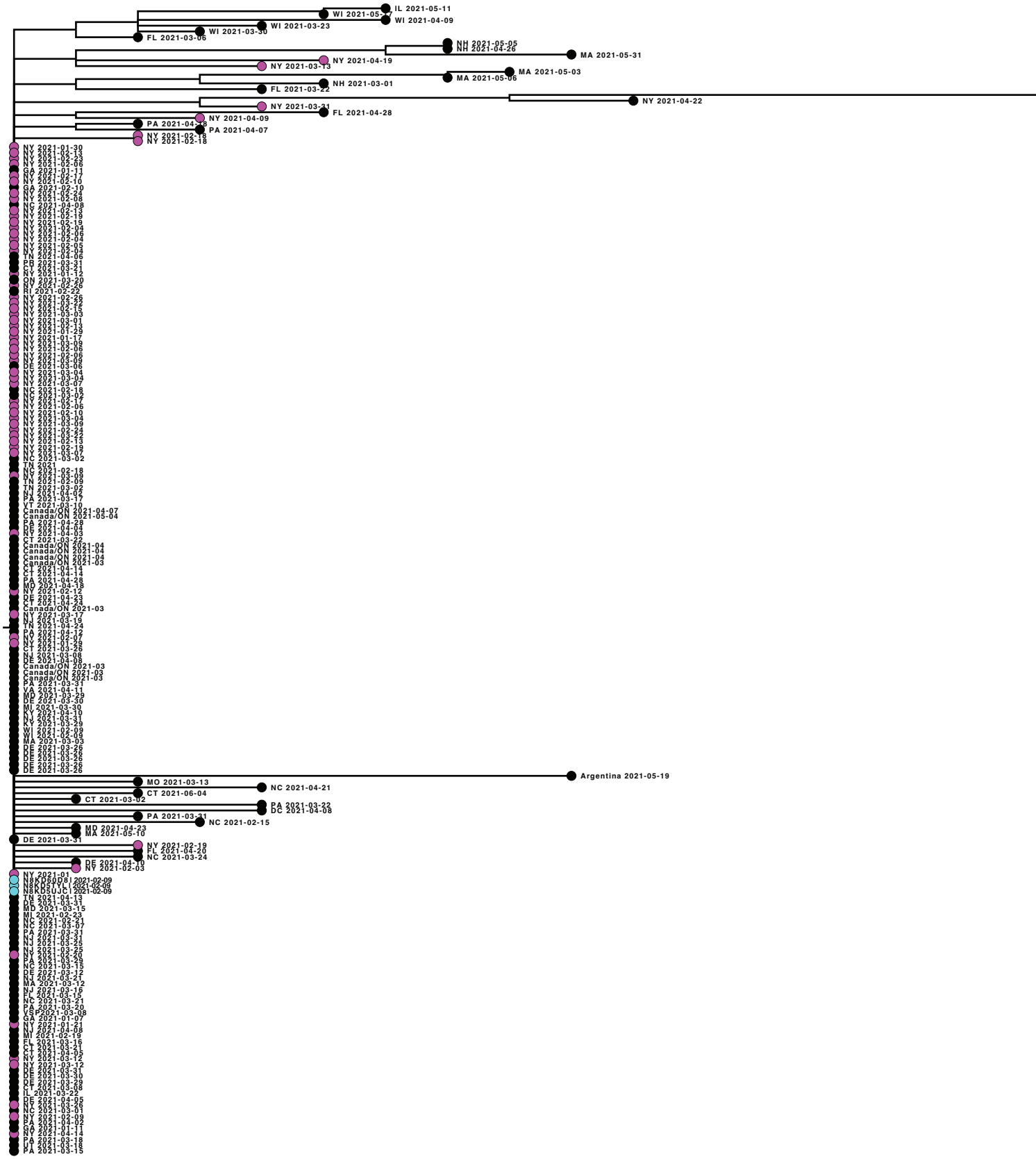

NY5
UT12
UT4
UT3
UT5
PR2
PR5
PR3
Frequency
1.00
0.75
0.50
0.25
0.00
C13372T
C6548T
T27447C
T7336C
A21366G
G20062T
T8644G
G5629T
A11201G
C9979T
N8KD4X9T
N8KD4Z1F
N8KECR4Q
N8KEDKLR
N8KCITCQ
N8KCOAXY
N8KCOZ3F
N8KCP3L3
N8KCO7GJ
N8KCOC6P
N8KCP0ZO
N8KCP10A
N8KCPJQH
N8KCPJQR
N8KCOXE3
N8KCOXG2
ZZYRBLBU
ZZYRBNJ3
ZZYRBSKN
ZZYRC7V3
ZZYRD912
ZZYRJLMR
N8KA7ZA8
N8KA7ZAK
N8KA7ZAQ
N8KA7ZAW

Supplemental Figure 1

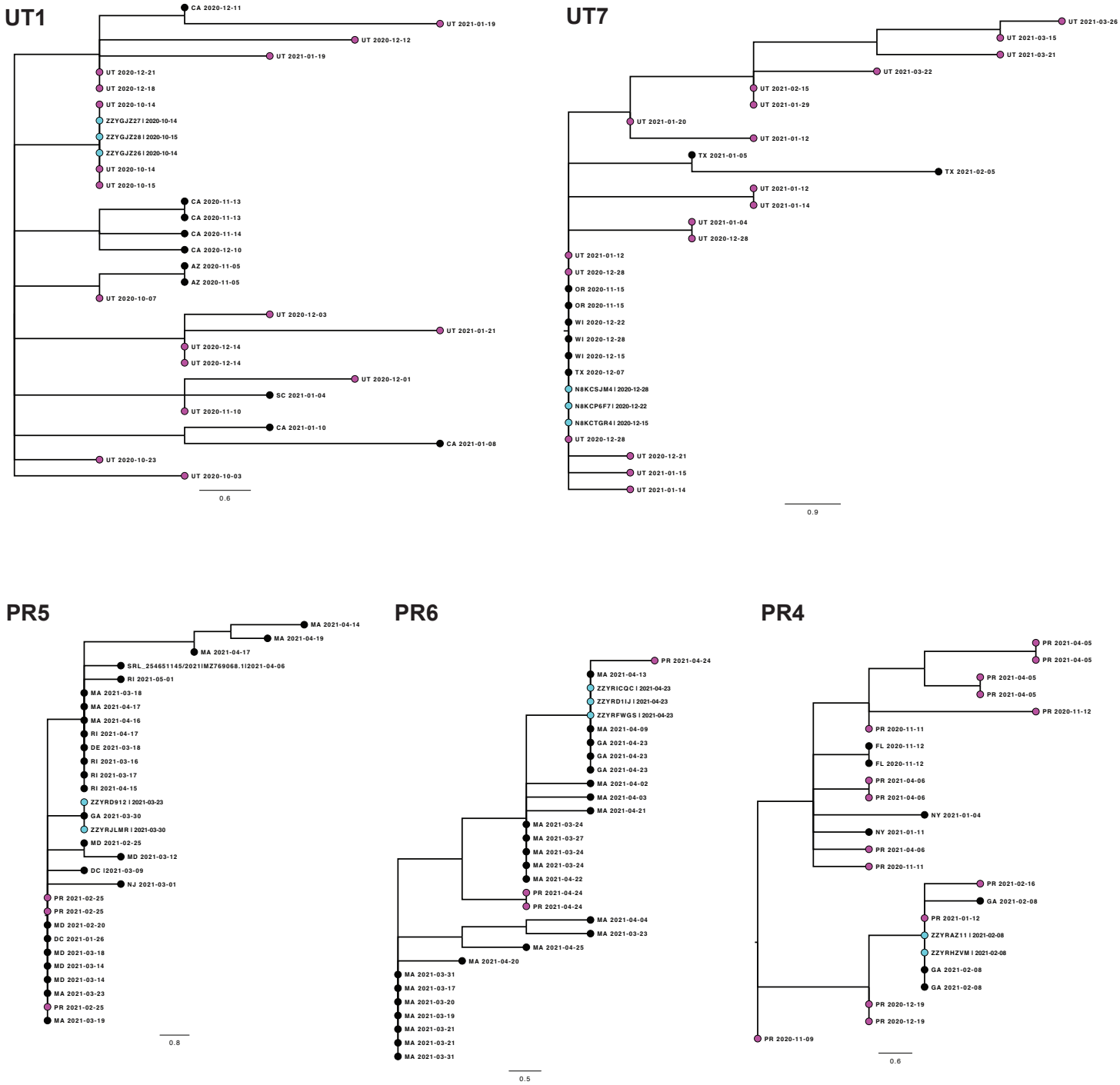


Supplemental Figure 2

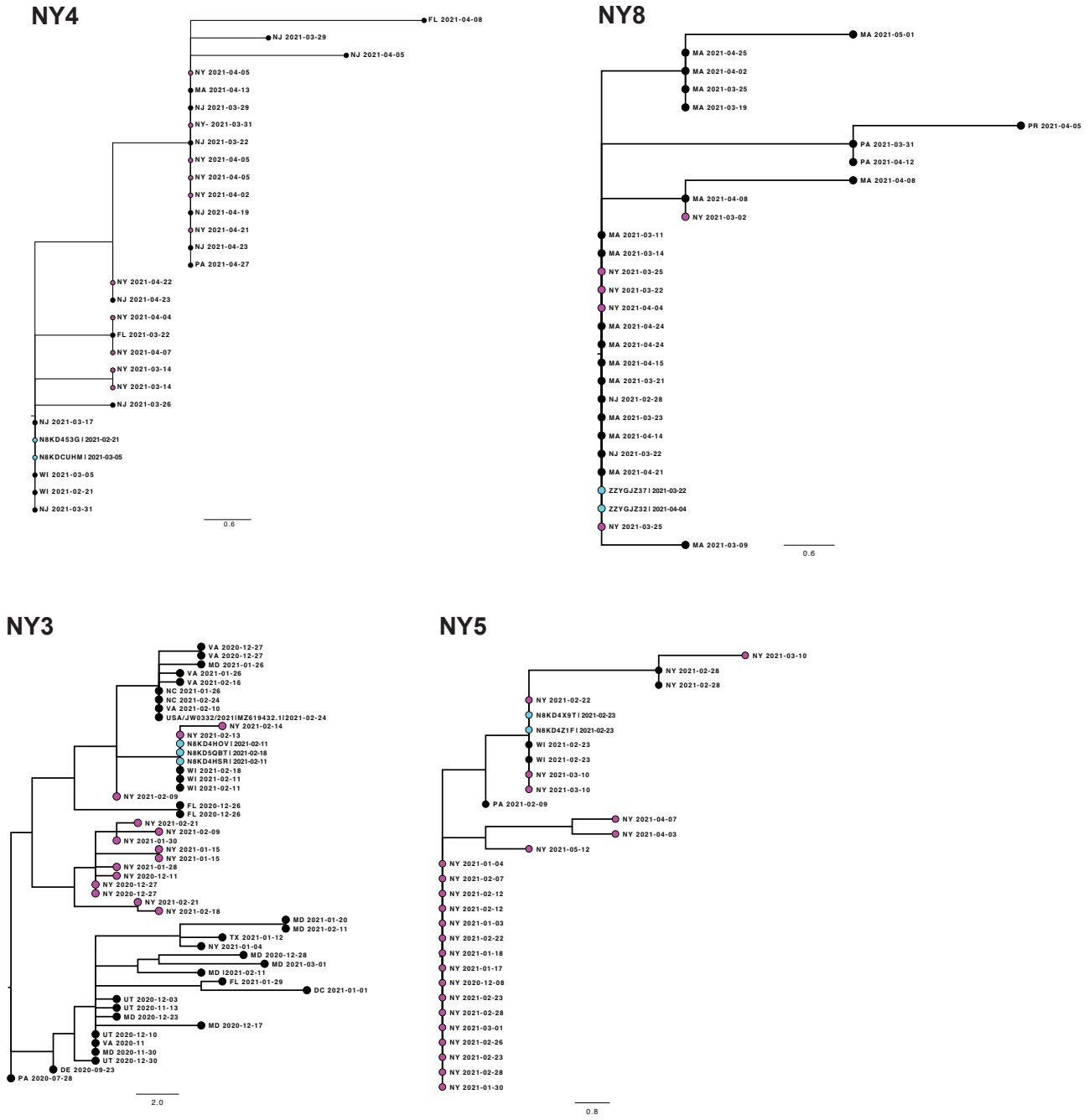
NY2



Supplemental Figure 3

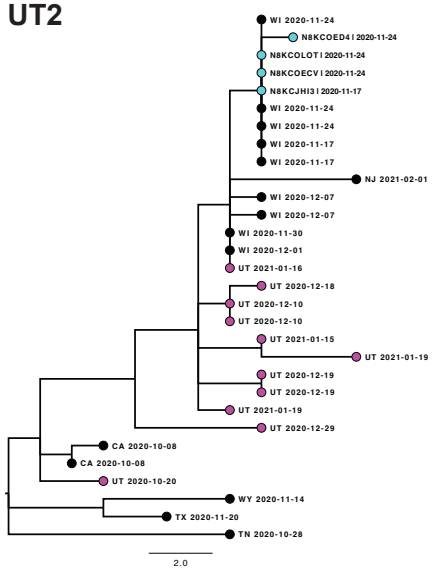


Supplemental Figure 4

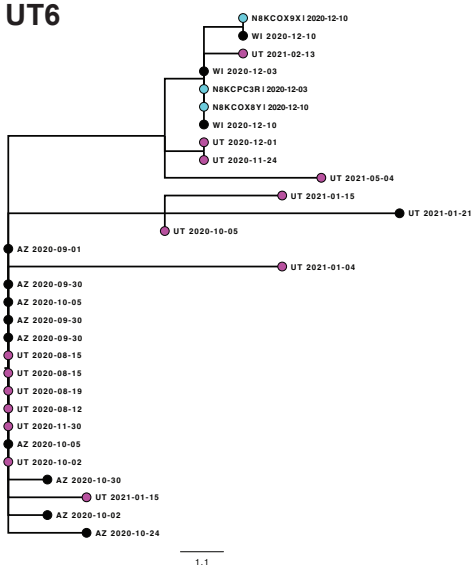


Supplemental Figure 5

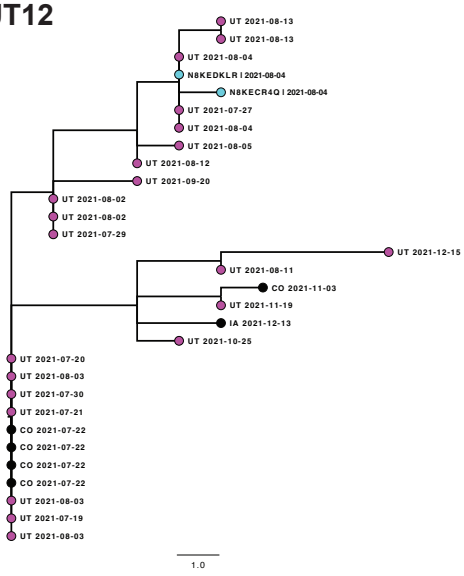
UT2



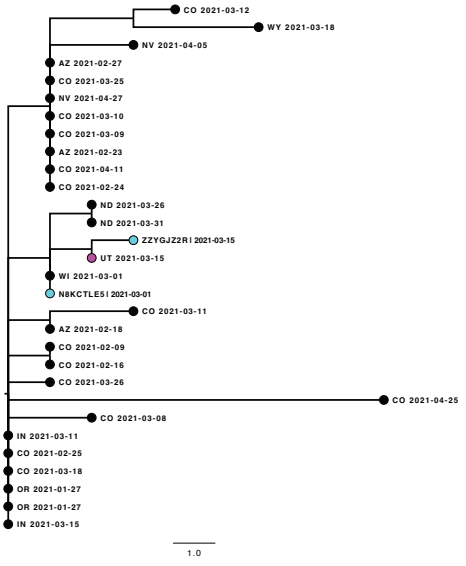
UT6



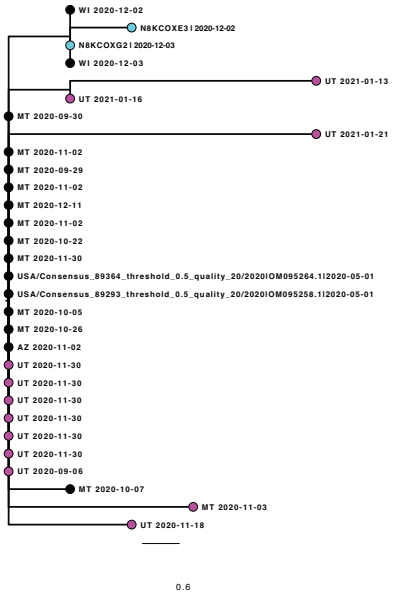
UT12



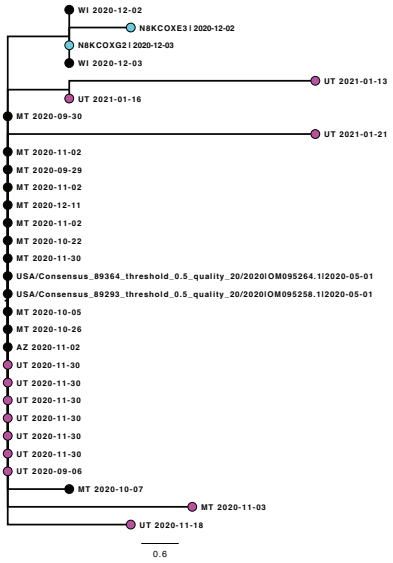
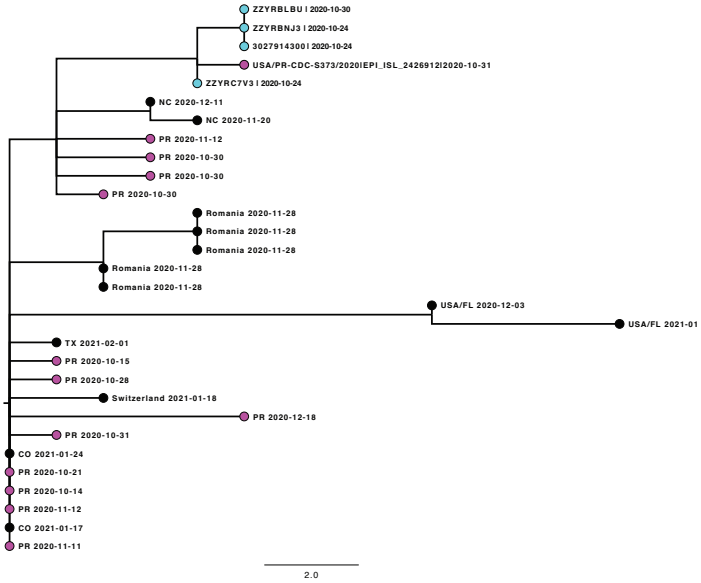
NY6



UT5



UT8



PR2