

City University of New York (CUNY)

CUNY Academic Works

Publications and Research

CUNY Graduate Center

2020

CoV Genome Tracker: tracing genomic footprints of Covid-19 pandemic

Saymon Akther
CUNY Graduate Center

Edgaras Bezrucenkovas
CUNY Hunter College

Brian Sulkow
CUNY Hunter College

Christopher Panlasigui
CUNY Hunter College

Li Li
CUNY Graduate Center

See next page for additional authors

[How does access to this work benefit you? Let us know!](#)

More information about this work at: https://academicworks.cuny.edu/gc_pubs/613

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).
Contact: AcademicWorks@cuny.edu

Authors

Saymon Akther, Edgaras Bezrucenkovas, Brian Sulkow, Christopher Panlasigui, Li Li, Wei-Gang Qiu, and Lia Di

CoV Genome Tracker: tracing genomic footprints of Covid-19 pandemic

Saymon Akther^{1,2}, Edgaras Bezrucenkovas², Brian Sulkow², Christopher Panlasigui², Li Li^{1,2},
Weigang Qiu^{1,2,3,*}, and Lia Di^{2,*}

¹Graduate Center, City University of New York, USA; ²Department of Biological Sciences, Hunter College, City University of New York, New York, New York 10065, USA; ³Department of Physiology and Biophysics & Institute for Computational Biomedicine, Weil Cornell Medical College, New York, New York 10021, USA

Emails: Saymon Akther <sakther@gradcenter.cuny.edu>; Edgaras Bezrucenkovas <edgaras993@gmail.com>; Brian Sulkow <drtwisto@gmail.com>; Christopher Panlasigui <christopher.panlasigui@gmail.com>; Li Li <lli4@gradcenter.cuny.edu>; *Co-correspondence: Weigang Qiu <weigang@genectr.hunter.cuny.edu> & Lia Di <dilie66@gmail.com>

Abstract

Summary: Genome sequences constitute the primary evidence on the origin and spread of the 2019-2020 Covid-19 pandemic. Rapid comparative analysis of coronavirus SARS-CoV-2 genomes is critical for disease control, outbreak forecasting, and developing clinical interventions. CoV Genome Tracker is a web portal dedicated to trace Covid-19 outbreaks in real time using a haplotype network, an accurate and scalable representation of genomic changes in a rapidly evolving population. We resolve the direction of mutations by using a bat-associated genome as outgroup. At a broader evolutionary time scale, a companion browser provides gene-by-gene and codon-by-codon evolutionary rates to facilitate the search for molecular targets of clinical interventions.

Availability and Implementation: CoV Genome Tracker is publicly available at <http://cov.genometracker.org> and updated weekly with the data downloaded from GISAID (<http://gisaid.org>). The website is implemented with a custom JavaScript script based on jQuery (<https://jquery.com>) and D3-force (<https://github.com/d3/d3-force>).

Contact: weigang@genectr.hunter.cuny.edu, City University of New York, Hunter College

Supplementary Information: All supporting scripts developed in JavaScript, Python, BASH, and PERL programming languages are available as Open Source at the GitHub repository <https://github.com/weigangq/cov-browser>.

31

32

Usages & Innovations

33

34

35

36

37

38

39

40

41

42

43

44

Genomic epidemiology comparatively analyzes pathogen genome sequences to uncover the evolutionary origin, trace the global spread, and reveal molecular mechanisms of infectious disease outbreaks including the latest coronavirus pandemic caused by the viral species SARS-CoV-2 (1–4). The unprecedented public-health crisis calls for real-time analysis and dissemination of genomic information on SARS-CoV-2 isolates accumulating rapidly in databases such as GISAID (<http://gisaid.org>) (5,6). To meet the challenge of real-time comparative analysis of SARS-CoV-2 genomes, we developed the CoV Genome Tracker (<http://genometracker.org>) with a supporting bioinformatics pipeline. Key features of the CoV Genome Tracker include interactive visualization and exploration of geographic origins, transmission routes, and viral genome changes of Covid-19 outbreaks (Fig 1). A companion comparative genomics website displays the 2003-2004 SARS-CoV and the 2019-2020 SARS-CoV-2 outbreaks in the evolutionary context of their wildlife relatives (1,7).

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

At the micro-evolutionary time scale, a key distinction of CoV Genome Tracker from the Nextstrain Covid-19 browser (<https://nextstrain.org/ncov>) (6) is our adoption of a haplotype network – instead of a phylogenetic tree – as the analytic framework as well as the visual guide (Fig 1). A haplotype network offers several advantages over a phylogenetic tree. First, at the time scale of days and months, loss and fixation of alleles are rare and the ancestral and descendant genotypes are both present in the population. As such, tree-based phylogenies can be misleading because tree-based phylogenetic algorithms compel all sampled genomes into leaf nodes regardless of ancestral or descendant genotypes, meanwhile introducing hypothetical ancestors as internal nodes. Second, phylogenetic reconstruction typically assumes a mutation-driven process with complete lineage sorting. Violation of these assumptions results in misleading evolutionary relations, for example, when recombination is present or when genes remain polymorphic (8,9). Third, a haplotype network requires less abstract comprehension of evolutionary processes than a phylogenetic tree does. For example, edges of a haplotype network depict genetic changes from a parent to a descendant genome, while branches of a phylogenetic tree represent genetic changes from a hypothetical ancestor to another hypothetical or sampled genome. Fourth, a haplotype network is more scalable than a phylogenetic tree as a visual tool. This is because the total number of nodes of a phylogenetic tree grows linearly with the number of genomes, resulting in a crowded visual space. In contrast, additional genomes add to the size

63 but not the total number of nodes of a haplotype network if they share the same haplotype se-
64 quence with previously sampled genomes.

65 A further innovation of the haplotype network used in the CoV Genome Tracker is the
66 inclusion of an outgroup genome to polarize all mutational changes. Conventional haplotype
67 networks show mutational differences but not mutational directions on edges (10–12). The di-
68 rected haplotype network in CoV Genome Tracker is thus informative for tracing the origin, fol-
69 lowing the spread, and forecasting the trend of Covid-19 outbreaks across the globe (Fig 1). To
70 date, one published study and two preprint manuscripts use haplotype networks to represent the
71 genealogy of SARS-CoV-2 isolates (13–15). These networks are however based on a much
72 smaller number of genomes, non-interactive, and non-directional.

73 At the macro-evolutionary time scale, CoV Genome Tracker provides more in-depth fea-
74 tures than the Nextstrain browser on SARS-CoV-2 genome evolution
75 (<https://nextstrain.org/groups/blab/sars-like-cov>) (6) (Supplemental Fig S1). Modeled after *Bor-*
76 *reliaBase* (<http://borreliabase.org>), a browser of Lyme disease pathogen genomes (16), the com-
77 parative genomics browser of CoV Genome Tracker provides analytical features including se-
78 quence alignments, gene trees, and codon-specific nucleotide substitution rates. As such, the
79 macro-evolutionary browser facilitates exploring the wildlife origin of SARS-CoV-2, identifying
80 functionally important gene sites based on sequence variability, and understanding mechanisms
81 of genome evolution including mutation, recombination and natural selection (3,4).

82 **Methods & Implementation**

83 The micro-evolutionary and macro-evolutionary browsers of the CoV Genome Tracker
84 are continuously updated according to the following workflows.

85 For the Covid-19 genome browser, we download genomic sequences and associated
86 metadata of SARS-CoV-2 isolates from GISAID (5), which are subsequently parsed with a PY-
87 THON script (“parse-metadata.ipynb”; all scripts available in GitHub repository <http://cov.ge->
88 [nometracker.org](http://cov.genometracker.org)). We use a custom BASH script (“align-genome.sh”) to align each genome to
89 an NCBI reference genome (isolate Wuhan-Hu-1, GenBank accession NC_045512) with Nu-
90 cmer4 (17), identify genome polymorphisms with Samtools and Bcftools (18), and create a hap-
91 lotype alignment using Bcftools. To minimize sequencing errors, we retain only phylogenetically
92 informative bi-allelic single-nucleotide polymorphism (SNP) sites where the minor-allele nucleo-
93 tide is present in two or more sampled genomes. To maximize network stability, a custom Perl

94 script (“impute-hap.pl”) is used to trim SNP sites at genome ends where missing bases are com-
95 mon, discard haplotypes with more than 10% missing bases, (optionally) impute missing bases
96 of a haplotype with homologous bases from a closest haplotype (19), and identify unique haplo-
97 types using the BioPerl package Bio::SimpleAlign (20). To root the haplotype network, we in-
98 clude the genome of a closely related bat isolate (RaTG13, GenBank accession MN996532) (1)
99 as the outgroup (using however only nucleotides at the SNP sites present among human iso-
100 lates).

101 We use two methods to infer a network genealogy of unique haplotypes. In one approach,
102 we infer a maximum parsimony tree using the DNAPARS program of the PHYLIP package (21).
103 A custom Perl script (“hapnet-pars.pl”) transforms the resulting maximum parsimony tree into a
104 phylogenetic network by replacing internal nodes with the nearest haplotypes where tree dis-
105 tances between the two are zero. Alternatively, we use a custom Perl script (“hapnet-mst.pl”) to
106 reconstruct a minimum-mutation network of unique haplotypes based on the Kruskal’s minimum
107 spanning tree (MST) algorithm implemented in the Perl module Graph ([https://metacpan.org/re-](https://metacpan.org/release/Graph)
108 [lease/Graph](https://metacpan.org/release/Graph)). Both Perl scripts polarize the edges of the haplotype network according to the
109 outgroup sequence by performing a depth-first search using the Perl module Graph::Tra-
110 versal::DFS (<https://metacpan.org/pod/Graph::Traversal::DFS>). The Perl scripts output a di-
111 rected graph file in the JavaScript Object Notation (JSON) format. The JSON network file is read
112 by a custom JavaScript, which layouts the website with the JavaScript library jQuery
113 (<http://jquery.com>) and creates an interactive force-directed rendering of the haplotype network
114 with the JavaScript library D3-force (<http://d3js.org>).

115 For the comparative genomics browser of CoV, we download genomes of a human-host
116 SARS-CoV-2 (isolate WIV2, GenBank accession MN996527), a human-host SARS-CoV (isolate
117 GD01, GenBank accession AY278489), and closely related coronavirus isolates from bat hosts
118 from the NCBI Nucleotide Database. We extract coding sequences from each genome and iden-
119 tify orthologous gene families using BLASTp (22). For each gene family, we obtain a codon
120 alignment using MUSCLE and Bioaln (23,24). We reconstruct maximum-likelihood trees for in-
121 dividual genes as well as for the whole genome based on a concatenated alignment of ten genes
122 using FastTree (25). For each gene, we estimate the maximum-parsimony number of nucleotide
123 changes at each codon position using DNACOMP of the PHYLIP package (21). Differences in
124 nucleotide substitution rates between the predominantly synonymous 3rd codon position and the
125 other two codon positions are indicative of forces of natural selection. For example, a higher
126 substitution rate at the 3rd codon position than the rate at the 1st and 2nd positions indicates

127 purifying selection while a higher or similar rate at the 1st and 2nd codon positions relative to the
128 rate at the 3rd codon position suggests adaptive diversification (e.g., at the Spike protein-encod-
129 ing locus) (2). The CoV comparative genomics browser is developed with the same software
130 infrastructure supporting *BorreliaBase* (<http://borreliabase.org>), a comparative genomics
131 browser of Lyme disease pathogens (16).

132 **Conclusion & Future Directions**

133 In summary, the CoV Genome Tracker facilitates up-to-date and interactive analysis of
134 viral genomic changes during current and future coronavirus outbreaks. The CoV Genome
135 Tracker uses a haplotype network, a more accurate and scalable model than a phylogenetic tree
136 to analyze and visualize genomic changes in the rapidly evolving SARS-CoV-2 population (6).
137 We improved upon conventional haplotype networks by resolving the direction of mutational
138 changes based on an outgroup genome (10,12). Future development will include implementing
139 probabilistic network algorithms such as maximum parsimony probability (10,11), developing
140 methods for testing network accuracy and stability, analyzing association between genomic
141 changes and network characteristics (e.g., association between the number of nonsynonymous
142 mutations and the in- and out-degrees of nodes), performance optimization, usability improve-
143 ments, and incorporating a mechanism for community feedback.

144 **Declaration**

145 **Availability of website & source codes**

146 CoV Genome Tracker is publically available at <http://cov.genometracker.org>. All source
147 codes are released as Open Source and available at <https://github.com/weigangq/cov-browser>.
148 The repository contains BASH, Perl, Python, R, and JavaScript codes for data processing pipe-
149 line, network reconstruction, and web development.

150 **Authors' Contributions**

151 S.A. implemented the genome processing pipeline and drafted the manuscript. E.B. de-
152 veloped the workflow for downloading and parsing data from the GISAID database. B.S. per-
153 formed network stability analysis and contributed to website design. C.P. prepared and main-
154 tains online documentation. L.L. contributed to network analysis, drafting manuscript, and online
155 documentation. W.Q. conceived the project, developed and implemented the network algorithm,
156 and drafted the manuscript. L.D. developed the meta-data pipeline, designed the website, im-
157 plemented JavaScript codes, and prepared the figures.

158 **Acknowledgements**

159 We gratefully acknowledge the authors, originating and submitting laboratories of the se-
160 quences from GISAID's EpiCoV™ Database on which this research is based. All submitters of
161 data may be contacted directly via www.gisaid.org. We thank Desiree Pante, Bing Wu, and Ra-
162 mandeep Singh for participation in data entry. We thank Dr Yozen Hernandez for system admin-
163 istration of computer networks. We thank Jonathan Sulkow for contributing to webpage design.

164 **Funding**

165 S.A. and L.L. are supported in part by the Graduate Program in Biology from the Graduate
166 Center, City University of New York. This work was supported in part by the National Institute of
167 Allergy and Infectious Diseases (NIAID) (AI139782 to W.Q.) of the National Institutes of Health
168 (NIH) of the United States of America. The funders had no role in study design, data collection
169 and analysis, decision to publish, or preparation of the manuscript.

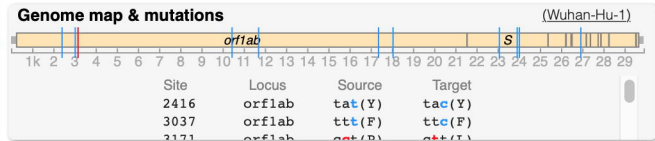
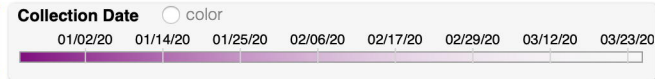
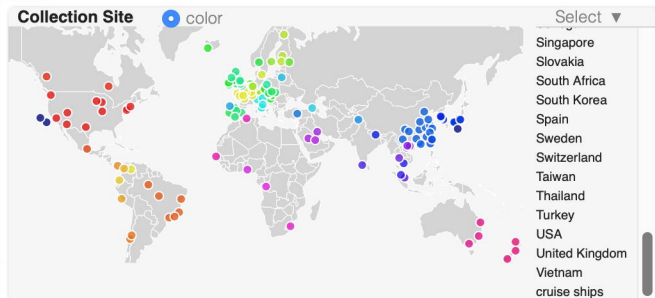
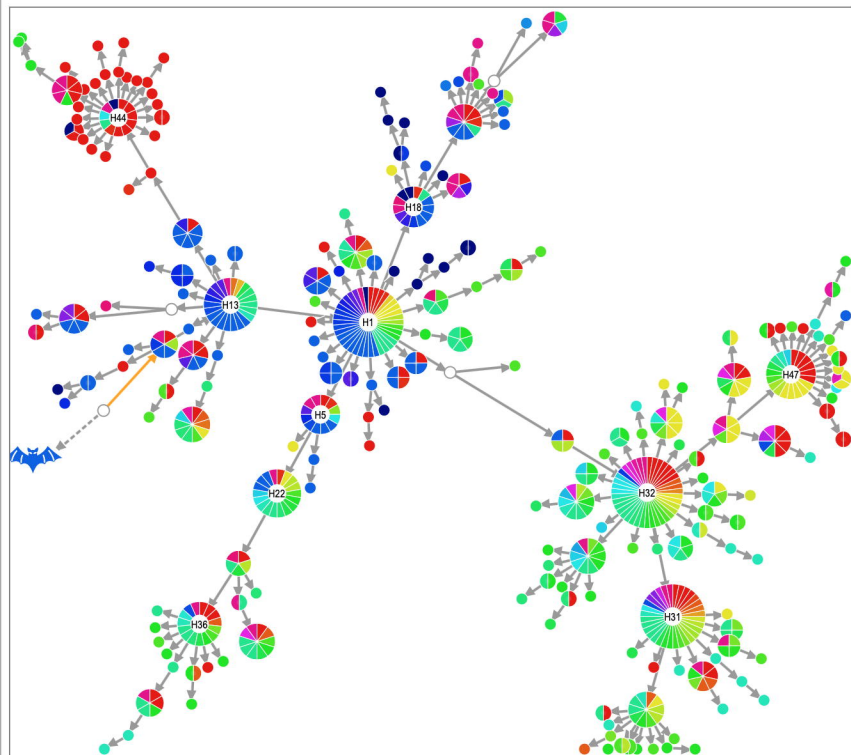
170 **References Cited**

- 171 1. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak as-
172 sociated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–3.
- 173 2. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiol-
174 ogy of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*
175 *Lond Engl*. 2020 22;395(10224):565–74.
- 176 3. Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, et al. Identifying SARS-
177 CoV-2 related coronaviruses in Malayan pangolins. *Nature*. 2020 Mar 26;1–6.
- 178 4. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of
179 SARS-CoV-2. *Nat Med*. 2020 Mar 17;1–3.
- 180 5. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to
181 reality. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull*. 2017 30;22(13).
- 182 6. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-
183 time tracking of pathogen evolution. *Bioinformatics*. 2018 Dec 1;34(23):4121–3.
- 184 7. Peeri NC, Shrestha N, Rahman MS, Zaki R, Tan Z, Bibi S, et al. The SARS, MERS and
185 novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats:
186 what lessons have we learned? *Int J Epidemiol*. 2020 Feb 22;
- 187 8. Edwards SV, Potter S, Schmitt CJ, Bragg JG, Moritz C. Reticulation, divergence, and the
188 phylogeography-phylogenetics continuum. *Proc Natl Acad Sci U S A*. 2016
189 19;113(29):8025–32.
- 190 9. Koch H, DeGiorgio M. Maximum Likelihood Estimation of Species Trees from Gene Trees
191 in the Presence of Ancestral Population Structure. *Genome Biol Evol*. 2020 Feb
192 1;12(2):3977–95.

- 193 10. Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealo-
194 gies. *Mol Ecol*. 2000 Oct;9(10):1657–9.
- 195 11. Leigh JW, Bryant D. popart: full-feature software for haplotype network construction.
196 *Methods Ecol Evol*. 2015;6(9):1110–6.
- 197 12. Múrias dos Santos A, Cabezas MP, Tavares AI, Xavier R, Branco M. tcsBU: a tool to ex-
198 tend TCS network layout and visualization. *Bioinformatics*. 2016 Feb 15;32(4):627–8.
- 199 13. Fang B, Liu L, Yu X, Li X, Ye G, Xu J, et al. Genome-wide data inferring the evolution and
200 population demography of the novel pneumonia coronavirus (SARS-CoV-2) [Internet].
201 *Evolutionary Biology*; 2020 Mar [cited 2020 Apr 5]. Available from: <http://bio->
202 rxiv.org/lookup/doi/10.1101/2020.03.04.976662
- 203 14. Yu W-B, Tang G, Zhang L, Corlett R. Decoding evolution and transmissions of novel
204 pneumonia coronavirus (SARS-CoV-2) using the whole genomic data [Internet]. 2020.
205 Available from: DOI: 10.12074/202002.00033
- 206 15. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-
207 2 genomes. *Proc Natl Acad Sci* [Internet]. 2020 Apr 8 [cited 2020 Apr 10]; Available from:
208 <https://www.pnas.org/content/early/2020/04/07/2004999117>
- 209 16. Di L, Pagan PE, Packer D, Martin CL, Akther S, Ramrattan G, et al. BorreliaBase: a phy-
210 logeny-centered browser of Borrelia genomes. *BMC Bioinformatics*. 2014 Jul 3;15(1):233.
- 211 17. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast
212 and versatile genome alignment system. *PLoS Comput Biol*. 2018;14(1):e1005944.
- 213 18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Align-
214 ment/Map format and SAMtools. *Bioinforma Oxf Engl*. 2009 Aug 15;25(16):2078–9.
- 215 19. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phyloge-
216 nies. *Mol Biol Evol*. 1999 Jan;16(1):37–48.
- 217 20. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl
218 toolkit: Perl modules for the life sciences. *Genome Res*. 2002 Oct;12(10):1611–8.
- 219 21. Felsenstein J. PHYLIP - Phylogeny Inference Package. *Cladistics*. 1989;5:164–6.
- 220 22. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
221 architecture and applications. *BMC Bioinformatics*. 2009;10:421.
- 222 23. Hernández Y, Bernstein R, Pagan P, Vargas L, McCaig W, Ramrattan G, et al. BpWrap-
223 per: BioPerl-based sequence and tree utilities for rapid prototyping of bioinformatics pipe-
224 lines. *BMC Bioinformatics*. 2018 Mar 2;19:76.
- 225 24. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high through-
226 put. *Nucleic Acids Res*. 2004;32(5):1792–7.
- 227 25. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for
228 large alignments. *PLoS One*. 2010;5(3):e9490.

230
231 **Fig 1. CoV Genome Tracker** uses a maximum-parsimony mutational network (*left panel*) to
232 represent genealogy of SARS-CoV-2 isolates during the 2019-2020 Covid-19 pandemic. The
233 network is interactively linked with geographic origins (color-coded, *top row, right*) and collection
234 dates (*2nd row, right*) of viral isolates, genomic locations (at n=146 SNP sites) and molecular
235 nature of mutations (*3rd row, right*), and isolate information searchable by GISAID accession (*4th*
236 *row, right*). Colored nodes represent haplotypes (n=212), a unique combination of nucleotides
237 at polymorphic genome positions. Open-circle nodes (n=4) represent hypothetical ancestors.
238 Each slice within a node, occupying one unit area, represents one or more viral isolates (n=2334
239 genomes downloaded from GISAID as of 3/29/2020) sharing a geographic origin. Thus, node
240 size is an indication of geographic diversity of a haplotype, not the number of isolates. In other
241 words, widely distributed genomes show as large nodes. Large nodes (containing >10 slices)
242 are labeled at the center. Each edge represents one or more mutational changes between a
243 parental and a descendant haplotype. Arrows indicate mutation directions determined according
244 to an outgroup genome (MN996532, strain “RaTG-13”, *bat icon*). The network is consistent with
245 a published one consisting of half the number of genomes (15). However, the maximum parsimony
246 network registers 59 (or 40.7%) sites that have changed more than once (i.e., homoplasy).
247 Causes of homoplasy include sequencing errors, presence of recombination, and the large evolutionary
248 distances between the outgroup and SARS-CoV-2 genomes. Nonetheless, CoV Genome Tracker
249 provides up-to-date genomic changes, helps trace the origin and spread, and
250 facilitates research into virulence mechanisms and clinical interventions on the current and future
251 coronavirus outbreaks.

252



Isolate EPI_ISL_ 🔍

Haplotype N=212
 a group of similar genomes

Isolate(s) N=2334 from [GISAID](#)
 from the same location

Mutation(s) at 146 genome sites
 genetic changes