Supplementary material

SARS-CoV-2 phylogeny during the early outbreak in the Basel area, Switzerland: import and spread dominated by a single B.1 lineage variant (C15324T)

Madlen Stange[1,2,3*], Alfredo Mari[1,2,3*], Tim Roloff[1,2,3*], Helena MB Seth-Smith[1,2,3*], Michael Schweitzer[1,2], Myrta Brunner[4], Karoline Leuzinger[5,6], Kirstine K. Søgaard[1,2], Alexander Gensch[1], Sarah Tschudin-Sutter[7], Simon Fuchs[8], Julia Bielicki[9], Hans Pargger[10], Martin Siegemund[10], Christian H Nickel[11], Roland Bingisser[11], Michael Osthoff[12], Stefano Bassetti[12], Rita Schneider-Sliwa[4], Manuel Battegay[7], Hans H Hirsch[5,6,7], Adrian Egli[1,2,+]

[1] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland
[2] Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland
[3] Swiss Institute for Bioinformatics, Basel, Switzerland
[4] Human Geography, University of Basel, Basel, Switzerland
[5] Clinical Virology, University Hospital Basel, Basel, Switzerland
[6] Transplantation & Clinical Virology, Department of Biomedicine, University of Basel, Basel, Switzerland
[7] Infectious Diseases and Hospital Epidemiology, University Hospital Basel and University of Basel, Basel, Switzerland
[8] Health Services for the City of Basel, Basel, Switzerland
[9] Pediatric Infectious Diseases, Children's University Hospital Basel, Basel, Switzerland
[10] Intensive Care Unit, University Hospital Basel, Basel, Switzerland
[11] Emergency Medicine, University Hospital Basel, Basel, Switzerland
[12] Internal Medicine, University Hospital Basel, Basel, Switzerland
*these four authors contributed equally to this work

+ correspondence
Adrian Egli, MD PhD
University Hospital Basel
Petersgraben 4
4031 Basel, Switzerland
Email: adrian.egli@usb.ch
Phone: +41 61 556 5749

Table of Contents

Supplementary methods

Patients, samples, and diagnosis

Respiratory samples from the University Hospital Basel and the University Children's Hospital Basel (UKBB) patients were tested for SARS-CoV-2: from January 23rd 2020 testing was based on current case definitions from the Federal Office of Public Health (FOPH); from 27th February additionally, all respiratory samples negative for other respiratory pathogens were tested. Patients samples which tested positive [1,2] for SARS-CoV-2 up to and including March 23rd were considered eligible for the present study. In total 6,943 diagnostic tests were performed during the study period. The 746 positively tested cases came predominantly from the administrative unit of Basel-City, Riehen, and Bettingen (418, 58%), while the remaining patients were from Basel-Landschaft and neighbouring cantons and countries.
For diagnosis, swabs from the naso- and oropharyngeal sites (NOPS) were taken, and combined into one universal transport medium tube (UTM, Copan). Total nucleic acids (TNAs) were extracted using the MagNA Pure 96 system and the DNA and viral RNA small volume kit (Roche Diagnostics, Rotkreuz, Switzerland) or using the Abbott m2000 Realtime System and the Abbott sample preparation system reagent kit (Abbott, Baar, Switzerland). Aliquots of extractions were sent for diagnosis to Charité, Berlin, Germany from January 23rd - 29th, and to Geneva to the National Reference Centre (NAVI) in Switzerland from January 29th. In-house analysis started February 27th as part of the hospital routine diagnostics as previously described [2].

Whole genome sequencing (WGS)

SARS-CoV-2 genomes were amplified following the amplicon sequencing strategy of the ARTIC protocol (https://artic.network/ncov-2019) with V.1 or V.3 primers [3]. In detail, real-time reverse transcriptase (RT) reactions were run to a total volume of 10µl extracted total nucleic acid. After some optimization, PCR used 25 cycles for samples with a diagnostic cycle threshold (C_t) value lower than 21 (viral loads higher than 8.2 log10 Geq/ml); 40 cycles for all other samples (lower viral load samples) and repeats. Purified amplicons were converted into Illumina libraries with Nextera Flex DNA library prep kit (Illumina) automated on a Hamilton STAR robot, using 5ng input DNA. 96 libraries were multiplexed and sequenced paired-end 150 nucleotides on an Illumina NextSeq 500 instrument.

Consensus sequence generation and detection of mutations

After demultiplexing using bcl2fastq software version v.2.17 (Illumina), COVGAP (COVid-19 Genome Analysis Pipeline) was used (**Figure S2**). This incorporates: quality filtering using trimmomatic software version v.0.38 [4] to remove Illumina adaptors and PCR primer sequences from read ends; removal of reads smaller than 127 bases, and removal of reads with a phred score under 20 (calculated across a 4-base sliding window). Quality filtered reads were mapped to the Wuhan-Hu-1 reference MN908947.3 [5] using the BWA aligner [6]. Reads flagged as mapping to the reference were retained [7], and are deposited under project PRJEB39887. SNPs and indels with respect to the reference sequence were called using pilon version 1.23 [8]. Pilon summary metrics 'alternative allele fraction' (AF) and 'depth of valid reads in pileup' (DP) were used to identify major and minor alleles across all bases, which is not implemented in pilon itself. Major alleles were called if supported by 70% of the reads covering the variant locus (AF) for any locus with a minimum of 50x coverage (DP). Variants were applied to the reference to produce a consensus sequence; any base position with less than 50x coverage was masked with ambiguous characters (Ns) using BCFTools version 1.10.2 [9]. Consensus sequences were accepted for further analysis when containing up to 10% Ns. Summary statistics, logs, coverage plots, and genome stack plots were generated using R version 3.6.0 and packages Gviz v1.30 [10], Sushi v1.23 [11], seqinr v3.6.1 [12], and ggplot2 v3.11 [13]. COVGAP also provides per genome quality control visual outputs (**Figure S5**) and is available at https://github.com/appliedmicrobiologyresearch.
Quality control statistics such as the relationship between C_t-value and number of mapped reads and coverage are presented in **Figure S3**. In general, we observed a negative trend linking C_t values and percentage of ambiguous bases (Ns) being called as a result of low coverage. Sequences passing the quality filter (n=533) showed a lower C_t value (median: 22.4±5.14) than the ones that failed (n =156; median: 35.75·9±5.75).

COVGAP Validation

We used a set of 15 randomly *in silico* mutated SARS-CoV-2 mock genomes for the validation of the specificity (identification of true negatives) and accuracy (identification of true negatives and true positives) of COVGAP. Additionally, the genome MT339040, which harbours an 81 nucleotide deletion in the ORF7a gene and a

further seven SNPs relative to the reference [14] was used. Together, the mock genomes possess 38 mutations including 30 SNPs, six deletions and two insertions across the reference genome MN908947.3. The genomes were then shredded to artificial paired-end 150 nucleotide reads using SAMtools wgsim [9] and processed by COVGAP. For validation purposes, original mock genomes and the COVGAP generated genomes from the shredded reads were aligned using Seaview v4.6 (Gouy, 2009) and clustalw (Sievers, 2011). A phylogeny was built using PhyML within Seaview with default parameters.

Phylogenetic lineage assignment of Basel samples

To assess the phylogenetic diversity of SARS-CoV-2 samples during the early phase of the pandemic we inferred the lineage assignment for each consensus sequence derived from the COVGAP pipeline using PANGOLIN ver. May 19th (Phylogenetic Assignment of Named Global Outbreak LINeages) [15] available at github.com/hCoV-2019/pangolin. Details on lineage summaries, describing which countries lineages have been reported from and where transmission events have been recorded, can be found at https://github.com/hCoV-2019/lineages. Lineage assignments were used to aid visualization of phylogenetic diversity in Basel in a global context. For global sequences we used the PANGOLIN lineage assignments as provided by GISAID (https://www.gisaid.org/; [16,17]) (details next section), which were used for plotting purposes on phylogenetic trees.
To compare the lineage diversity in Switzerland and Basel-City to neighbouring European countries (Austria, France, Germany, and Italy) during the early phase of the pandemic, we visualized relative abundances of lineages using all high-quality, on GISAID (downloaded June 22nd, 2020) available consensus sequences for the time until March 23rd from Austria (N = 188), France (N = 230), Germany (N = 133), and Italy (N = 98). For Switzerland (N = 673), we combined our sequences (N = 468) with other sequence data from Switzerland published on GISAID. To infer the diversity for canton Basel-City (including Bettingen and Riehen) excluding sequences that were obtained from commuters, we used the Basel-City portion (N = 376) of the Basel area cohort excluding samples from patients from cities outside of the administrative district of Basel-City. We calculated Simpson diversity (inverse Simpson concentration) as implemented in the SpadeR package v.0.0.1 [18-20], which controls for lineage abundance differences between the countries, which is dependent on available sequence data, and which ranges from 0 (no diversity) to indefinite (large diversity).

Analysing Basel SARS-CoV-2 genomes in global phylogenetic context

High-quality and full-length consensus sequences and corresponding metadata (sample ID, date of sample, geographic location of sampling, PANGOLIN lineage) from global viruses were downloaded from GISAID on June 22nd, 2020, making 49,284 individual genome sequences. 43,252 sequences were retained after filtering for genomes with under 10% ambiguous characters (Ns) (author Genivaldo Gueiros Z. Silva) [21]. Metadata and consensus sequences of the Basel samples and global data from GISAID were combined for further joint analysis, which were performed using custom R scripts and the nextstrain command line interface analysis pipeline v.2.0.0 (nextstrain.org) and augur v.8.0.0 [22].
Dates in our study samples correspond to date of sampling. Sequences were filtered by date from December 1st 2019 to March 23rd 2020 using an R custom script in R version 4.0.0 [23] and packages tidyr ver.1.1.0 [24], dplyr ver. 1.0.0. [25], and readr ver. 1.3.1. [26]: 15,973 consensus sequences, including the Basel area sequences, remained. These time-filtered sequences were sub-sampled by geographic location to 30 sequences per country and month. Non-human derived viruses as well as sequences with other ambiguous characters (Us), as well as those from cruise ships, and duplicated sequences defined by the nextstrain team as of June 24th (https://github.com/nextstrain/ncov/) were excluded using *augur filter* [22] resulting in 2,485 sequences for the final phylogenetic analysis dataset.
Consensus sequences were aligned to the NCBI Refseq sequence Wuhan-Hu-1 reference MN908947.3 using mafft v7.467 with method FFT-NS-fragment [27] and options --reorder --keeplength --mapout --kimura 1 --addfragments --auto. The resulting alignment was end-trimmed to remove low-quality bases (bases 1-55; 29804-29903). We masked homoplasic sites (**Table S2**) that have no phylogenetic signal [28] (deposited at https://github.com/W-L/ProblematicSites_SARS-CoV2). Given the constant updates which this homoplasic site list undergoes, we used to the data released on June 19th, 2020. Masking was done using *augur mask*.
The resulting alignment was analysed in IQ-TREE 2 [29] for tree inference using *augur tree* with substitution model GTR+G. The tree in Newick format was then subjected, together with the date information of each genome and the initial sequence alignment, to an estimation of the evolutionary rate by a regression of the divergence (number of mutations) against the sampling date using TreeTime [30] implemented in *augur refine*. Genomes or branches that deviated more than four interquartile ranges from the root to the tip versus the time tree were removed as likely outliers. The resulting time-calibrated and divergence trees were re-rooted to

MN908947.3 and MT291826.1, the first official cases and published genomes of SARS-CoV-2 from Wuhan, China.
Ancestral trait reconstruction of each patient's viral genome was done for region (continent) and country as well as region and country of exposure using *augur* traits with a sampling bias correction of 2.5. Internal nodes and tips (actual genomes) were annotated regarding their nucleotide and amino acid changes in relation to the reference using *augur ancestral* and *augur translate*, respectively. All data were exported as json files (supplementary files) using *augur export v2* to be visualized in *auspice* v2 [22].
Identified clades of interest were further inspected for existing epidemiological links using data collected by the University Hospital.

Identifying genomes belonging to GISAID emerging clade A20/15324T

To identify a possible geographic origin of the synonymous C15324T mutation in *ORF1ab*, we performed a search on all available GISAID genomes as of August 12th, 2020. We downloaded all high quality and complete genomes that were assigned do GISAID legacy clade G (corresponds to clade 20A) and PANGOLIN lineage B.1 (all three are mostly congruent [31]) with a collection date between December 2019 and March 23rd, 2020 (N = 2,856). We used Nextclade version 0.3.5 (https://clades.nextstrain.org) to infer genomic mutations and filtered for sequences that contained C15324T. This procedure allowed avoidance of homoplasic mutations at this site. Further, we downloaded metadata for all high quality and complete genomes (as of August 12th, 2020) irrespective of clade to calculate summary statistics of number of genomes sequenced per country.

Identification of S-gene D614G mutation in Basel sequences

We screened the early phase Basel sequences for the mutation at nucleotide position 23,403 based on the alignment to the ***Wuhan-Hu-1*** reference sequence MN908947.3. Viral load (C_t-value) of patients that carried lineages with a mutated S-D614G gene (N = 274) were compared to patients that carried the ancestral allele (N = 12) using a Mann-Whitney U test.

Supplementary Results

Basel samples in phylogenetic global context continued

Cluster B.1.5

Isolates that are assigned to lineage B.1.5 make up 2·6% of USB isolates (**Figure S6B**). They all share the A20268G mutation. Three unresolved branches defined by at least one additional mutation each, diverge from the internal node consisting of, from top to bottom, six (C25658T), six (C28854T), and one (G25483A, C4893T, C23380A, C26509T [mutations in order of temporal appearance]) Basel area isolates. Individual isolates can exhibit one to three additional mutations. Isolates date from March 13th to March 23rd with an inferred node age of February 19th (CI: January 13th-Ferbuary 20th, 2020). No social connections for transmission patterns within each branch could be inferred from the available patient data. Searching the clade defining mutations in the nextstrain.org phylogeny we gain the following insights. Mutation C25658T (plus the clade defining A20268G) is found in one isolate (Oman/RESP-20-6701/2020 from March 28th); C28854T is found 17 isolates, two of which show no additional mutations (Norway/2088/2020 from March 17th, Latvia/045/2020 from March 22nd) just like two of our isolates (42193056, 42189239). Derived isolates originate from Switzerland, Scotland, Romania, USA, Taiwan, and England. Mutation G25483A recorded in a single isolate (42202280) is not currently reported in the nextstrain.org phylogeny.

Cluster B.1.8

Isolates that are assigned to lineage B.1.8 make up 0·7% of USB isolates (**Figure S6C**). They all share the A24862G mutation. Isolates date from March 14th to March 22nd with an inferred internal node age of February 1st (CI: January 12th-March 8th, 2020). Two isolates (42191012, 42202147) exhibit the identical mutational pattern (additional T658C, C28829T) but have no known epidemiological link. Our own global comparison identified an isolate from Germany (Germany/NRW-34/2020 from March 16th) that exhibits the same mutations. Searching the clade defining mutation in the nextstrain.org phylogeny does not yield better insights into the evolution of the lineages as no isolates with the same pattern could be identified.

Family clusters within lineage B.2

We identified eight genomes that were assigned to lineage B.2 (**Figure S6D**). They all share the G26144T mutation that translates into amino acid change ORF3a-G251V and date from March 13th to March 22nd with an inferred internal node age of January 15th (CI: January 13th-January 18th, 2020). This cluster harbours two household transmission clusters: *Family 2* with two members and *Family 3* with three members. These two clusters share C14805T (synonymous in *ORF1ab*) and exhibit unique additional mutations C9319T (synonymous in *ORF1ab*) and G12278T (ORF1ab-A4005S), G26730T (M-V70F), G29414T (N-A381S), respectively. We find no evidence of further community transmission. These mutational combinations are not currently represented in the full global phylogeny (nextstrain.org), suggesting that quarantine measures were effective in these cases and inhibited further transmission events.

Supplementary Tables

Table S1. Counts and description of the in silico mutated genome community used for COVGAP validation. Each observation consists of the genome position multiplied by the number of samples in which it appears. Attached as additional file.

Table S2. Nucleotide position in relation to the Wuhan-Hu1 reference sequence that were masked for phylogenetic inferences, due to homoplasies. Inferred by contributors to https://github.com/W-L/ProblematicSites_SARS-CoV2**.**

Start position	End position
635	635
2091	2091
2094	2094
3145	3145
3564	3564
4050	4050
5736	5736
6869	6869
8022	8022
8790	8790
10129	10129
11074	11074
11083	11083
11535	11535
13402	13402
13408	13408
13476	13476
13571	13571
14277	14277
15922	15922
16887	16887
19484	19484
21575	21575
22335	22335
24389	24389
24390	24390
24933	24933
26549	26549
29037	29037
29553	29553

Table S3. List of countries that recorded genomes with mutation C15324T and number of total genomes sequenced until March 23rd 2020.

Country	number genomes with C15324T	Total genomes sequenced until March 23rd	% genomes with mutation	% of population sequenced	Population
Argentina	1	4	25·00	0·00001	45,195,774
Australia	14	1092	1·28	0·00428	25,499,884
Austria	3	244	1·23	0·00271	9,006,398
Belgium	40	268	14·93	0·00231	11,589,623
Benin	1	6	16·67	0·00005	12,123,200
Bosnia and Herzegovina	2	12	16·67	0·00037	3,280,819
Brazil	1	226	0·44	0·00011	212,559,417
Canada	7	405	1·73	0·00107	37,742,154
Chile	2	120	1·67	0·00063	19,116,201
Costa Rica	1	40	2·50	0·00079	5,094,118
Democratic Republic of the Congo	11	35	31·43	0·00004	89,561,403
England	4	5643 (UK)	0·01	0·00831	67,886,011
France	69	369	18·70	0·00057	65,273,511
Germany	2	147	1·36	0·00018	83,783,942
Hungary	2	18	11·11	0·00019	9,660,351
Iceland	5	522	0·96	0·15297	341,243
India	1	119	0·84	0·00001	1,380,004,385
Israel	1	72	1·39	0·00083	8,655,535
Japan	3	343	0·87	0·00027	126,476,461
Luxembourg	24	116	20·69	0·01853	625,978
Morocco	3	13	23·08	0·00004	36,910,560
Netherlands	2	617	0·32	0·00360	17,134,872
Oman	1	21	4·76	0·00041	5,106,626
Portugal	8	570	1·40	0·00559	10,196,709
Russia	1	59	1·69	0·00004	145,934,462
Scotland	4	5643 (UK)	0·01	0·00831	67,886,011
Senegal	3	24	12·50	0·00014	16,743,927
South Korea	1	196	0·51	0·00038	51,269,185
Switzerland	57 (386)*	213 (675)*	26·8 (57·2)*	0·00780	8,654,622
Taiwan	3	95	3·16	0·00040	23,816,775
USA	1	4150	0·02	0·00125	331,002,651
Vietnam	1	59	1·69	0·00006	97,338,579

**** Number in brackets summarize counts of genomes from GISAID plus genomes from this study***

Table S4. GISAID identifiers and dates of sampling for all sequences that belong to emerging clade 20A/15324T with a collection date until March 23rd, 2020 (N = 279). Supplied as additional file.

Table S5. Diversity indices for SARS-CoV-2 lineages in Switzerland and neighbouring countries.

Country	Coefficient of co-variation	Shannon Entropy H'	Shannon Diversity	Simpson Concentration Index D'	Simpson Diversity
Austria	1·577	1·680	5·365	0·2679	3·7322
France	1·762	0·421	1·524	0·8144	1·2278
Germany	1·358	1·637	5·137	0·2583	3·8715
Italy	0·751	1·067	2·908	0·3971	2·5181
Switzerland	2·768	0·869	2·385	0·6156	1·6243

References

1. Goldenberger D, Leuzinger K, Sogaard KK, et al. Brief validation of the novel GeneXpert Xpress SARS-CoV-2 PCR assay. *J Virol Methods* 2020; **284**: 113925.
2. Leuzinger K, Roloff T, Gosert R, et al. Epidemiology of SARS-CoV-2 Emergence Amidst Community-Acquired Respiratory Viruses. *J Infect Dis* 2020.
3. Quick J. nCoV-2019 sequencing protocol. protocols.io 2020. dx.doi.org/10.17504/protocols.io.bdp7i5rn.
4. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**(15): 2114-20.
5. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020; **579**(7798): 265-9.
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**(14): 1754-60.
7. Wala J, Zhang CZ, Meyerson M, Beroukhim R. VariantBam: filtering and profiling of next-generational sequencing data using region-specific rules. *Bioinformatics* 2016; **32**(13): 2029-31.
8. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014; **9**(11): e112963.
9. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011; **27**(21): 2987-93.
10. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods Mol Biol* 2016; **1418**: 335-51.
11. Phanstiel DH, Boyle AP, Araya CL, Snyder MP. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* 2014; **30**(19): 2808-10.
12. D. C, J.R. L. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. . In: U. B, M. P, H.E. R, M. V, eds. Structural Approaches to Sequence Evolution Biological and Medical Physics, Biomedical Engineering. Berlin, Heidelberg: Springer; 2007.
13. Wickham H. ggplot2: Elegant Graphics for Data Analysis. https://ggplot2-book.org/.
14. Holland LA, Kaelin EA, Maqsood R, et al. An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). *J Virol* 2020; **94**(14).
15. Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020.
16. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 2017; **1**(1): 33-46.
17. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* 2017; **22**(13): 30494.
18. Chao A, Chiu CH, Hsieh TC. Proposing a resolution to debates on diversity partitioning. *Ecology* 2012; **93**(9): 2037-51.
19. Chao A, Jost L. Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution* 2015; **6**(8): 873-82.
20. Chao A, Wang Y, Jost L. Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods Ecol Evol* 2013; **4**: 1091-100.
21. Cock PJ, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009; **25**(11): 1422-3.
22. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018; **34**(23): 4121-3.
23. Team RC. R: A language and environment for statistical computing. https://www.r-project.org/.
24. Wickham H, Henry L. tidyr: Tidy Messy Data. R package version 1.1.0. . 2020. https://CRAN.R-project.org/package=tidyr.
25. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.0. . 2020. https://CRAN.R-project.org/package=dplyr.
26. Wickham H, J H, Francois R. readr: Read Rectangular Text Data. R package version 1.3.1. . 2018. https://CRAN.R-project.org/package=readr.
27. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013; **30**(4): 772-80.
28. De Maio N, C W, N G. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/10 (accessed May 29th 2020.

29. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**(5): 1530-4.
30. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* 2018; **4**(1): vex042.
31. GISAID - Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses. https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/ (accessed 21.08.2020 2020).

Supplementary Figures Legends

Figure S1. Age distribution by sex for the time period between February 24th and March 23rd for all tests, positive tests, and for patient isolates from which whole genomes were generated. Solid lines for females, dashed lines for males.

Figure S2. The COVGAP pipeline. The steps shown ensure the calculation of high quality consensus sequences. Particularly, information on read coverage is retained and used both in the variant calling procedure and in the draft of the consensus independently from the called variants. Finally, the quality of the genome from each sample is scored by %Ns, which determines whether the produced sequence is retained or discarded.

Figure S3. COVGAP evaluation of sequencing quality parameters.
Of the original 746 samples, 689 successfully sequenced. Number of mapped reads across all SARS-CoV-2 positive samples successfully sequenced from 26th of February till 23th of March: (n=689), of which 533 passed the quality filter, and 156 failed. 468 of the samples passing the quality filters were matching the cohort eligibility criteria and therefore were further described in the present study. **A.** Number of mapped reads against Ct values from diagnostic tests; **B.** Number of mapped reads against percentage of Ns in the consensus.

Figure S4. COVGAP identifies all SNPs in the mock genomes. A. Levenshtein distance between the mutations in the input mock genomes (y axis) and in the genomes recovered by COVGAP (x axis); the marginal plots show the frequency of presence / detection across samples. Only two deletions (5845, 16281) and one insertion (16145) were not detected. **B**. Phylogeny of input and COVGAP-derived consensus output genomes, showing that all SNPs were identified.

Figure S5. Representative diagnostic output from COVGAP. This output generated per sample, indicates (from top to bottom): **A**. the coverage –not represented if over 1000x; **B**. which variants were detected in which position of the genome, and their corresponding annotation; **C**. low coverage regions (under 50x); **D**. genome annotation; and **E**. genome size markers as reference. Of note, a report generated in parallel provides further information on the variants, including which amino acids are affected by the variant.

Figure S6. Divergence tree and zoom into additional sequence clusters, which did not result in large community spread. A. Isolates from Basel area cohort in global context. **B**. A small clade assigned to B.1.5 consists of two clusters with an accumulation of samples from Basel. **C**. Cluster within lineage B.1.8 with two Basel samples without known epidemiological link. **D**. Two family cluster within lineage B.2 that did not spread further into the Basel community.

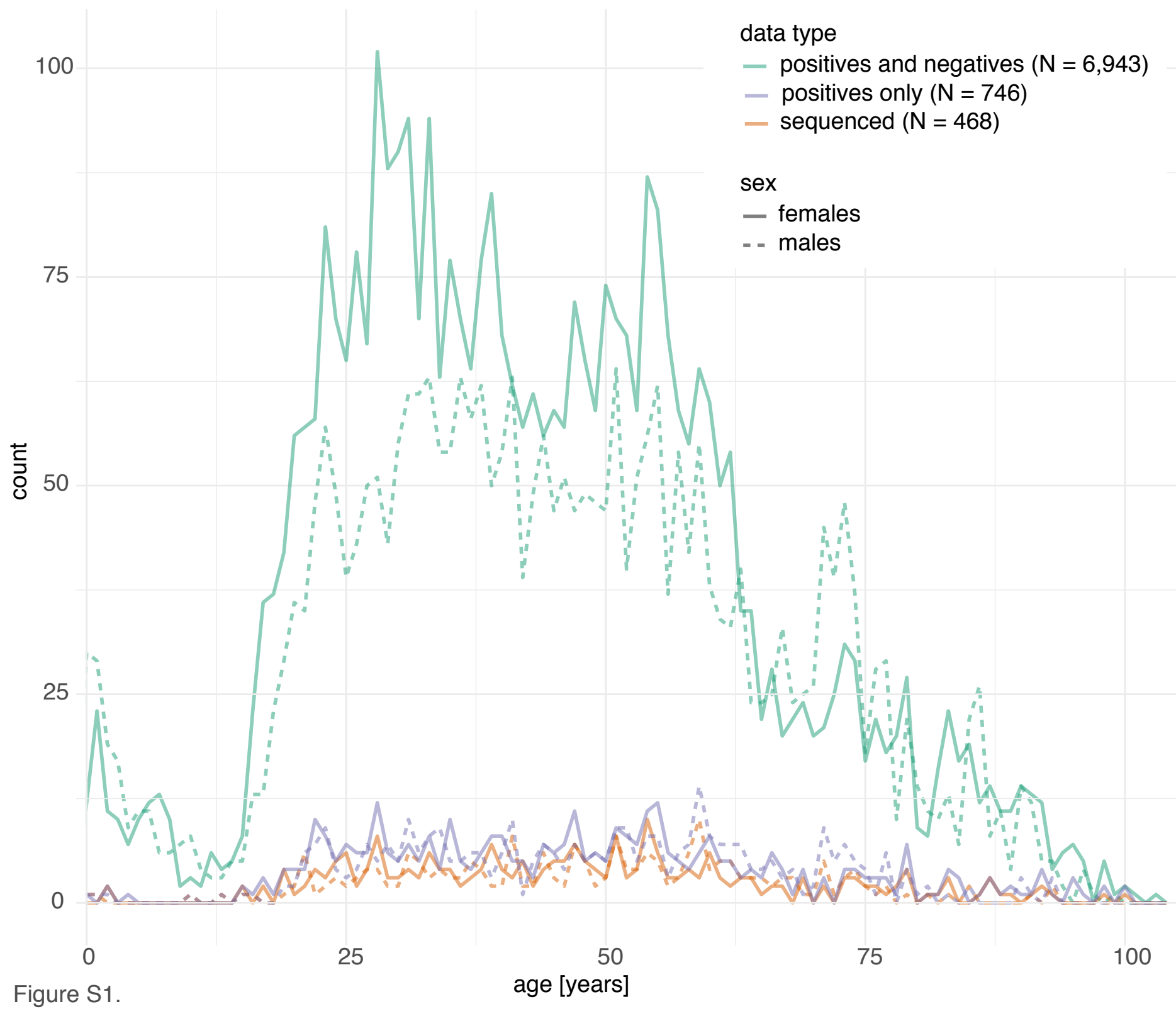
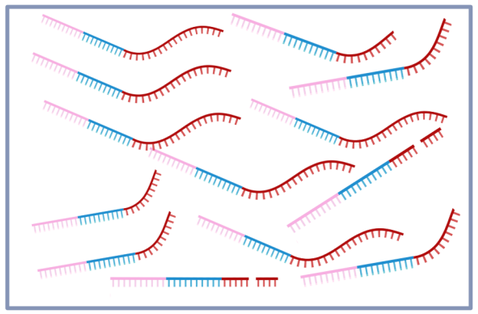

data type
positives and negatives (N = 6,943)
positives only (N = 746)
sequenced (N = 468)
sex
females
males
count
100
75
50
25
0
0
25
50
75
100
age [years]

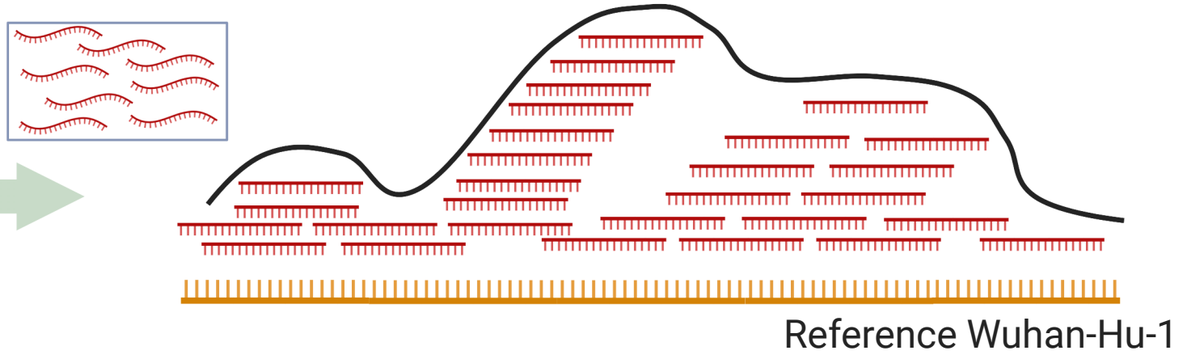
Figure S1.

Quality filter:



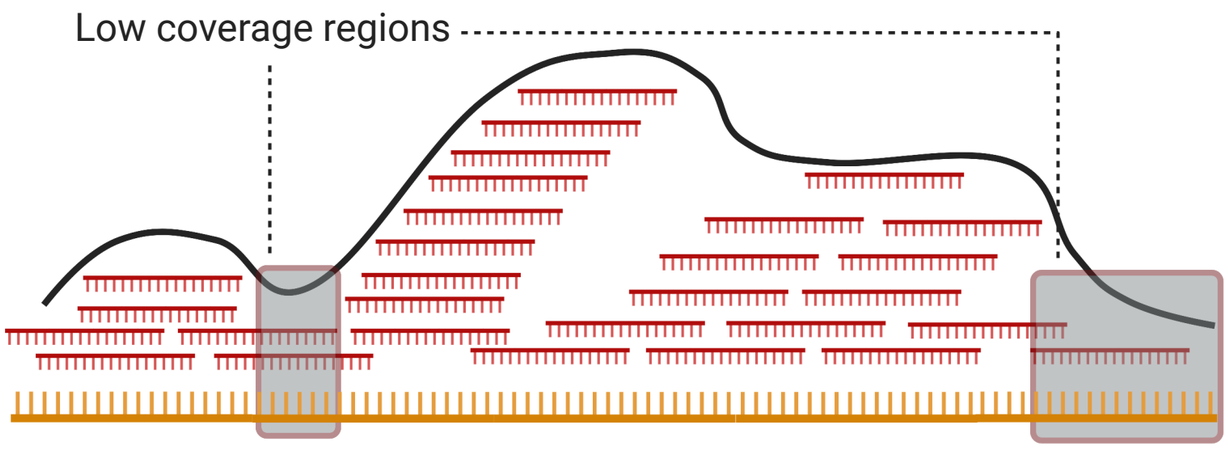
Adapter trimming
Primer trimming
Length filter
Phred score filter
Alignment to the reference
Mapping reads only retained

Mapping:



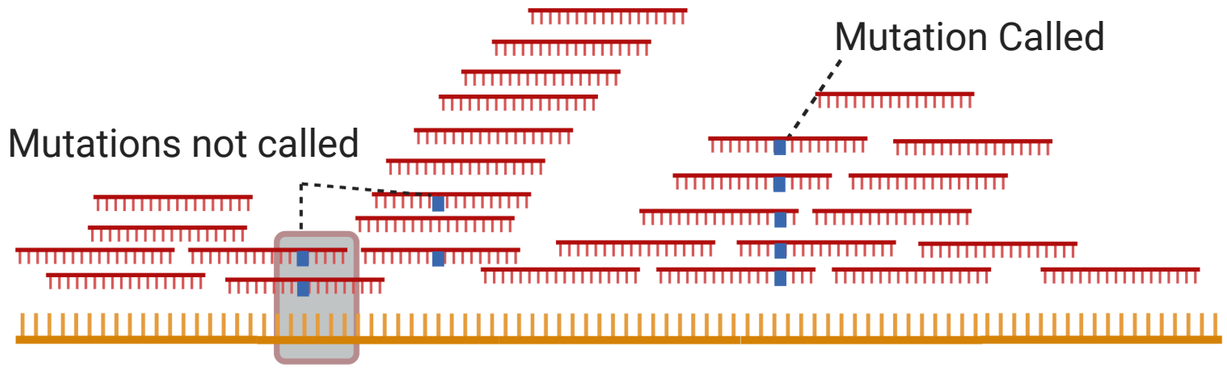
Coverage filter :

only positions covered by more than 50 reads are considered highly covered

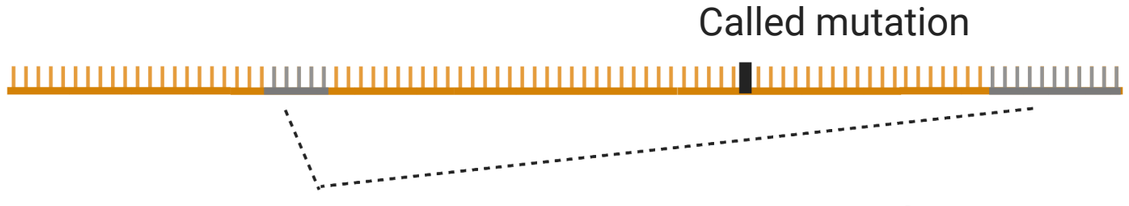


Calling only highly read-supported variants:

only mutations supported by 70% of available reads and not in low covered regions are retained



Consensus sequence generation:



QC checking:

sequences with > 10% Ns are discarded

QC checking:

sequences with < 10% Ns are retained

Figure S2.

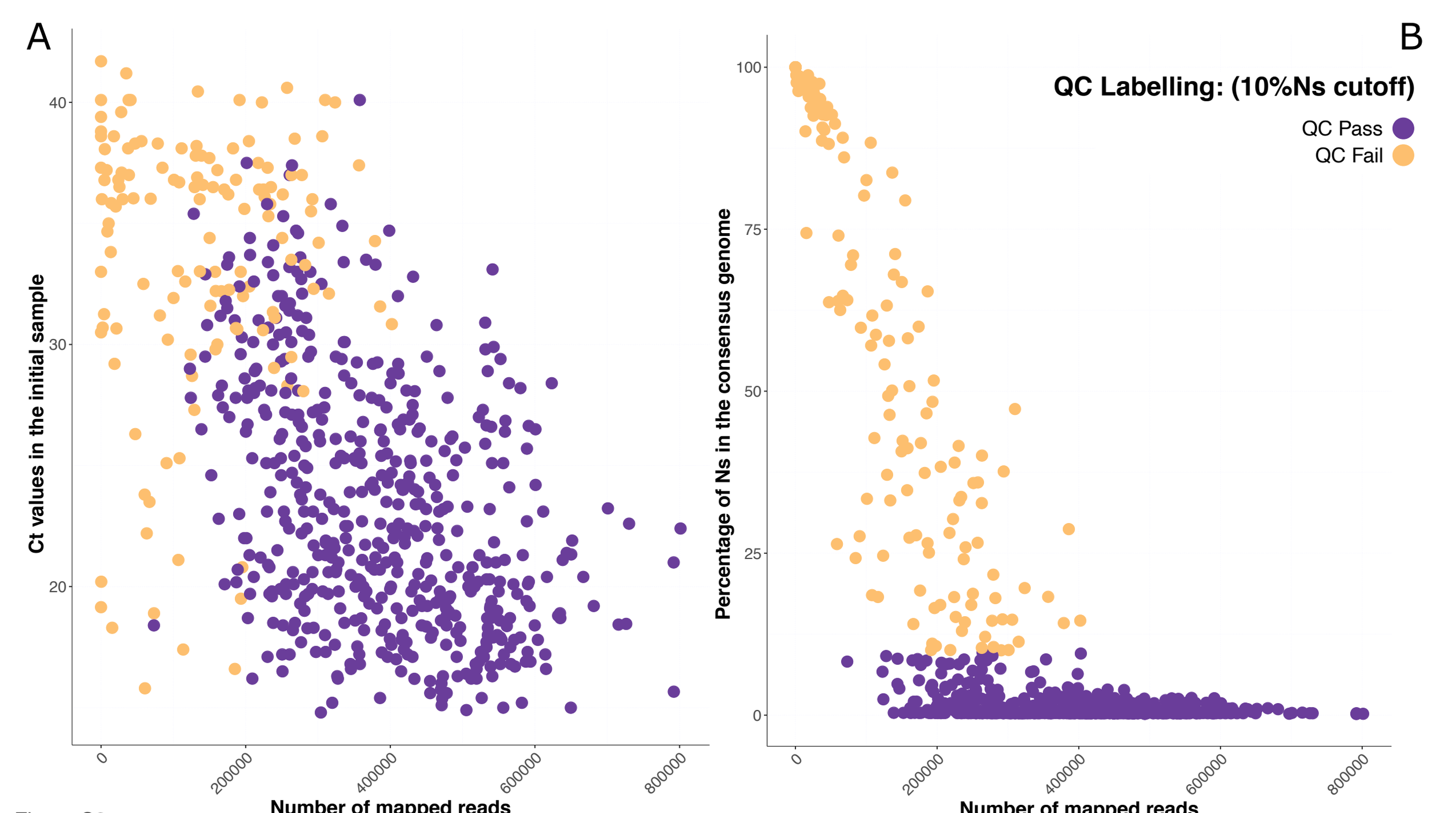
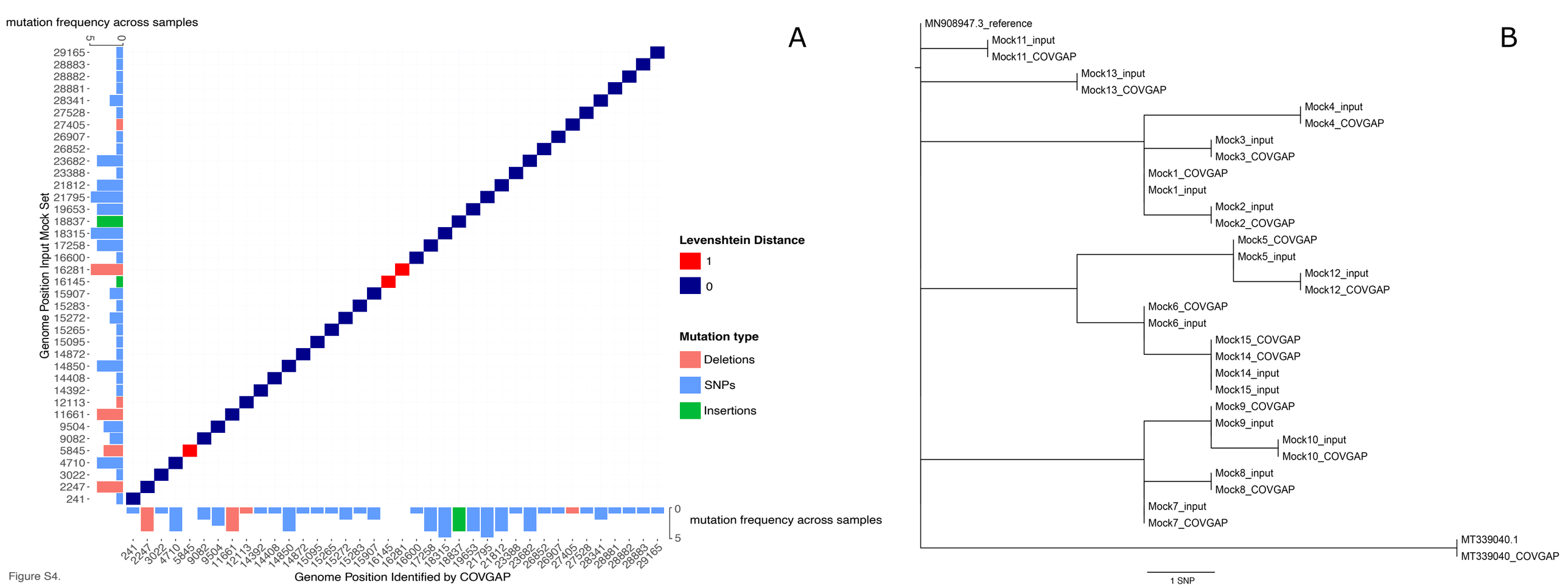

A
B
Ct values in the initial sample
Percentage of Ns in the consensus genome
Number of mapped reads
Number of mapped reads
QC Labelling: (10%Ns cutoff)
QC Pass
QC Fail
40
30
20
100
75
50
25
0
0
200000
400000
600000
800000

Figure S3.


A
mutation frequency across samples
5
0
Genome Position Input Mock Set
29165
28883
28882
28881
28341
27528
27405
26907
26852
23682
23388
21812
21795
19653
18837
18315
17258
16600
16281
16145
15907
15283
15272
15265
15095
14872
14850
14408
14392
12113
11661
9504
9082
5845
4710
3022
2247
241
Levenshtein Distance
1
0
Mutation type
Deletions
SNPs
Insertions
0
5
mutation frequency across samples
241
2247
3022
4710
5845
9082
9504
11661
12113
14392
14408
14850
14872
15095
15265
15272
15283
15907
16145
16281
16600
17258
18315
18837
19653
21795
21812
23388
23682
26852
26907
27405
27528
28341
28881
28882
28883
29165
Genome Position Identified by COVGAP
Figure S4.
B
MN908947.3_reference
Mock11_input
Mock11_COVGAP
Mock13_input
Mock13_COVGAP
Mock4_input
Mock4_COVGAP
Mock3_input
Mock3_COVGAP
Mock1_COVGAP
Mock1_input
Mock2_input
Mock2_COVGAP
Mock5_COVGAP
Mock5_input
Mock12_input
Mock12_COVGAP
Mock6_COVGAP
Mock6_input
Mock15_COVGAP
Mock14_COVGAP
Mock14_input
Mock15_input
Mock9_COVGAP
Mock9_input
Mock10_input
Mock10_COVGAP
Mock8_input
Mock8_COVGAP
Mock7_input
Mock7_COVGAP
MT339040.1
MT339040_COVGAP
1 SNP

Sample number: 42219995

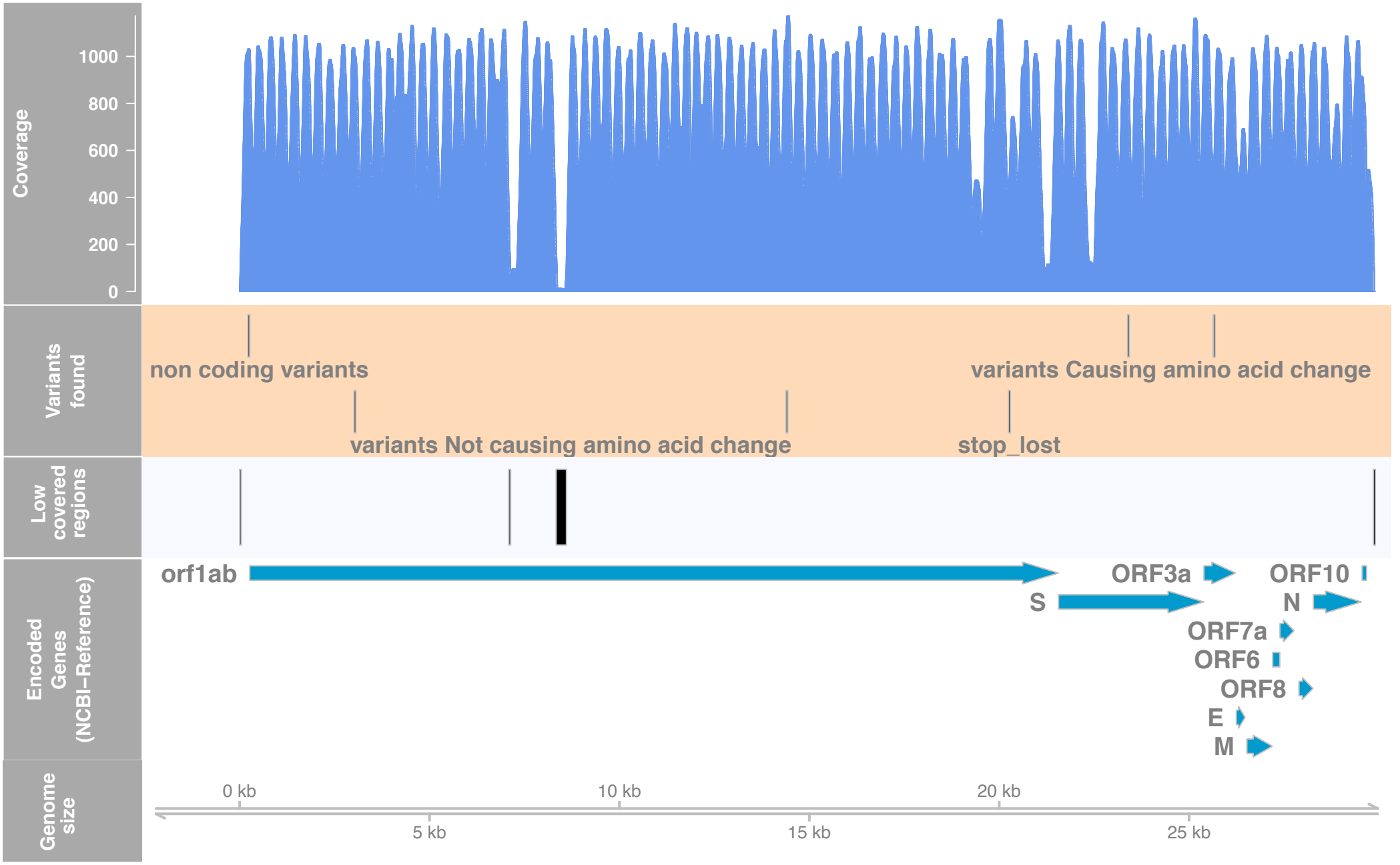


Figure S5.

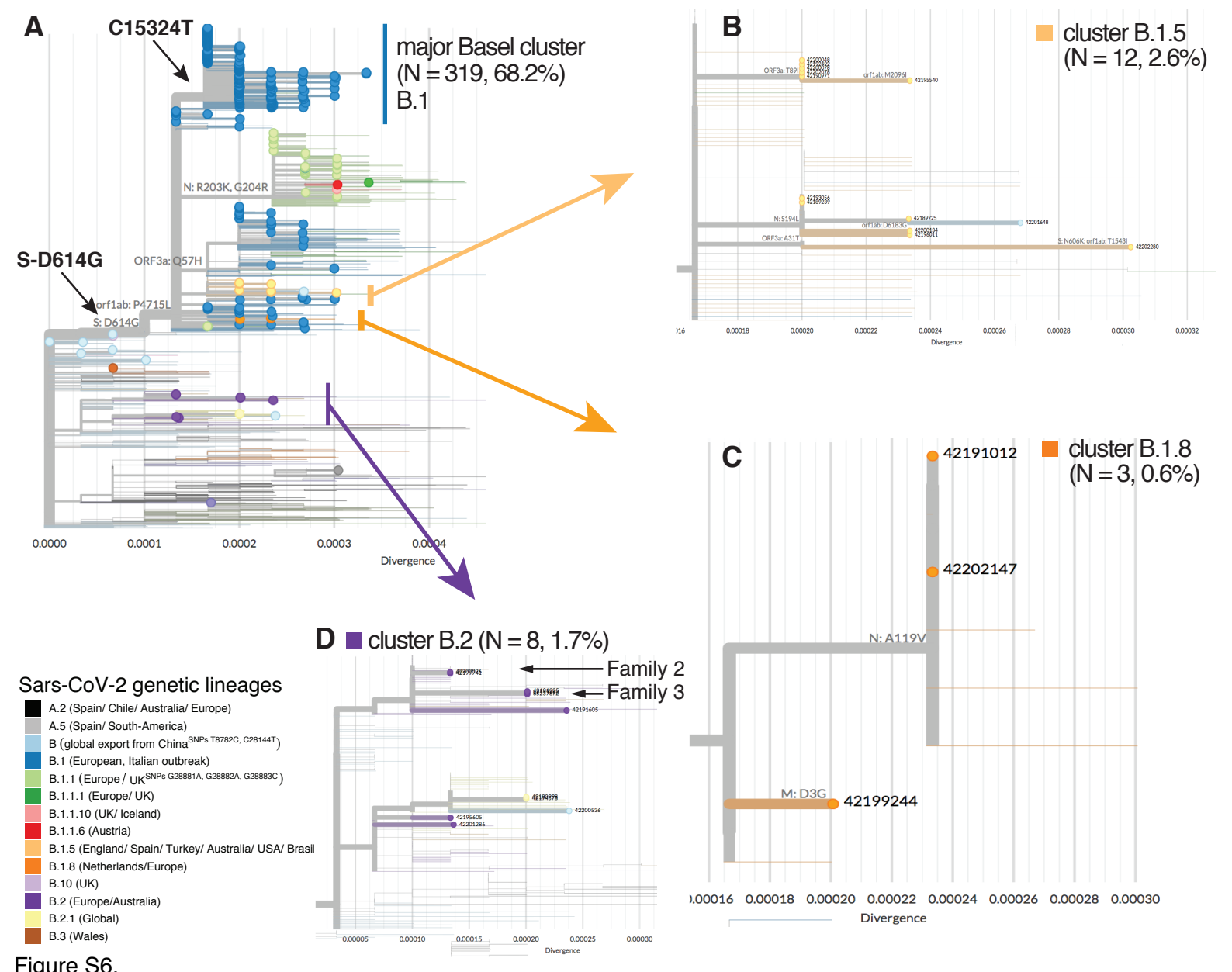

A
C15324T
major Basel cluster
(N = 319, 68.2%)
B.1
N: R203K, G204R
S-D614G
ORF3a: Q57H
orf1ab: P4715L
S: D614G
0.0000
0.0001
0.0002
0.0003
0.0004
Divergence
B
cluster B.1.5
(N = 12, 2.6%)
ORF3a: T89I
orf1ab: M2096I
42195540
N: S194L
orf1ab: D6183G
42189725
42201648
ORF3a: A31T
S: N606K; orf1ab: T1543I
42202280
0.00018
0.00020
0.00022
0.00024
0.00026
0.00028
0.00030
0.00032
Divergence
C
cluster B.1.8
(N = 3, 0.6%)
42191012
42202147
N: A119V
M: D3G
42199244
0.00016
0.00018
0.00020
0.00022
0.00024
0.00026
0.00028
0.00030
Divergence
D
cluster B.2 (N = 8, 1.7%)
Family 2
Family 3
42191605
42200536
42195605
42201286
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
Divergence
Sars-CoV-2 genetic lineages
A.2 (Spain/ Chile/ Australia/ Europe)
A.5 (Spain/ South-America)
B (global export from China SNPs T8782C, C28144T)
B.1 (European, Italian outbreak)
B.1.1 (Europe / UK SNPs G28881A, G28882A, G28883C)
B.1.1.1 (Europe/ UK)
B.1.1.10 (UK/ Iceland)
B.1.1.6 (Austria)
B.1.5 (England/ Spain/ Turkey/ Australia/ USA/ Brasil
B.1.8 (Netherlands/Europe)
B.10 (UK)
B.2 (Europe/Australia)
B.2.1 (Global)
B.3 (Wales)

Figure S6.