Within-patient genetic diversity of SARS-CoV-2

Jack Kuipers[1,2,*], Aashil A Batavia[1,2,4*], Kim Philipp Jablonski[1,2,*], Fritz Bayer[1,2], Nico Borgsmüller[1,2], Arthur Dondi[1,2], Monica-Andreea Drăgan[1,2], Pedro Ferreira[1,2], Katharina Jahn[1,2], Lisa Lamberti[1,2], Martin Pirkl[1,2], Susana Posada-Céspedes[1,2], Ivan Topolsky[1,2], Ina Nissen[1], Natascha Santacroce[1], Elodie Burcklen[1], Tobias Schär[1], Vincenzo Capece[1], Christiane Beckmann[3], Olivier Kobel[3], Christoph Noppen[3], Maurice Redondo[3], Sarah Nadeau[1,2], Sophie Seidel[1,2], Noemie Santamaria de Souza[5], Christian Beisel[1], Tanja Stadler[1,2], Niko Beerenwinkel[1,2,**]

1 Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
2 SIB Swiss Institute of Bioinformatics, Basel, Switzerland
3 Viollier AG, Allschwil, Switzerland
4 Department of Pathology and Molecular Pathology, USZ, Zurich, Switzerland
5 Institute of Microbiology, D-BIOL, ETH Zurich, Zurich, Switzerland

* Equal contributions
** Corresponding author

Abstract

SARS-CoV-2, the virus responsible for the current COVID-19 pandemic, is evolving into different genetic variants by accumulating mutations as it spreads globally. In addition to this diversity of consensus genomes across patients, RNA viruses can also display genetic diversity within individual hosts, and co-existing viral variants may affect disease progression and the success of medical interventions. To systematically examine the intra-patient genetic diversity of SARS-CoV-2, we processed a large cohort of 3939 publicly-available deeply sequenced genomes with specialised bioinformatics software, along with 749 recently sequenced samples from Switzerland. We found that the distribution of diversity across patients and across genomic loci is very unbalanced with a minority of hosts and positions accounting for much of the diversity. For example, the D614G variant in the Spike gene, which is present in the consensus sequences of 67.4% of patients, is also highly diverse within hosts, with 29.7% of the public cohort being affected by this coexistence and exhibiting different variants. We also investigated the impact of several technical and epidemiological parameters on genetic heterogeneity and found that age, which is known to be correlated with poor disease outcomes, is a significant predictor of viral genetic diversity.

Author Summary

Since it arose in late 2019, the new coronavirus (SARS-CoV-2) behind the COVID-19 pandemic has mutated and evolved during its global spread. Individual patients may host different versions, or variants, of the virus, hallmarked by different mutations. We examine the diversity of genetic variants coexisting within patients across a cohort of 3939 publicly accessible samples and 749 recently sequenced samples from Switzerland. We find that a small number of patients carry most of the diversity, and that patients with more diversity tend to be older. We also find that most of the diversity is concentrated in certain regions and positions of the virus genome. In particular, we find that a variant reported to increase infectivity is among the most diverse positions. Our study provides a large-scale survey of within-patient diversity of the SARS-CoV-2 genome.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the cause of COVID-19, spread globally from its origins in Wuhan, China, toward the end of 2019, resulting in the World Health Organisation declaring COVID-19 a pandemic in March 2020. Initially diagnosed as a pneumonia of unknown origin, huge research efforts have drastically developed our clinical knowledge of COVID-19 concerning its etiology [1], mode of transmission [2], identification of individuals at risk [3] and potential treatment strategies [4].

COVID-19 is a respiratory disease marked by a variety of symptoms including a persistent cough, fatigue and anosmia [5], however, a subpopulation of SARS-CoV-2-infected individuals remains asymptomatic [6]. The large variation in the severity of experienced symptoms, along with the virus incubation time ranging from 5-14 days [7,8], may have contributed to its rapid propagation. At the end of September 2020, there have been more than 34 million confirmed cases with over a million deaths worldwide [9].

SARS-CoV-2 belongs to the family *Coronaviridae* and is the latest of three zoonotic coronaviruses that have spilled over to infect humans in the past two decades following SARS-CoV in 2003 and MERS-CoV in 2012 [10]. The SARS-CoV-2 single-stranded RNA genome consists of at least 13 open reading frames (ORFs) spanning 29,903 nucleotides [11]. ORF1ab codes for a polyprotein from which the non-structural proteins originate, including the RNA-dependent RNA polymerase; this is then followed by the Spike (S) gene, the Envelope (E) gene, the Matrix (M) gene, the Nucleocapsid (N) gene and a host of accessory genes [12]. Much like the related SARS-CoV, the Spike protein present on the surface of the SARS-CoV-2 virion is central for its access to the target host cell via binding of the hACE2 receptor, S protein priming and finally fusion of the viral and target cell membranes [12–14]. Zeigler et al. identified co-expression of hACE2 and TMPRSS2 (a host protease required for viral entry) in lung type II pneumocytes, nasal goblet secretory cells, and ileal absorptive enterocytes; these are currently thought to be the target cells of SARS-CoV-2 leading to COVID-19 [15].

Viral isolation and sequencing efforts throughout the world have permitted the interrogation of the SARS-CoV-2 genome, providing a deeper understanding of the evolution of the virus, its proximal origin and revealing patterns of its global spread [16,17]. These efforts are largely centered on reverse transcription of the viral RNA genome followed by PCR amplicon and hybrid capture based sequencing using Oxford Nanopore and Illumina sequencers [18]. The sequences generated are largely deposited in data repositories such as the sequence read archive (SRA) and GISAID [19]. The availability of these sequences has allowed for platforms such as Nextstrain [20] to continually update and track viral evolution and spread, and to illuminate the viral genetic diversity among carriers [21–24].

At this inter-host level, epidemiological studies often employ per-patient consensus sequences, which summarize each patient's virus population into a single sequence and ignore minor variants. However, RNA viruses generally exhibit high mutation rates due to error-prone viral RNA polymerases, which typically leads to the presence of various viral variants within a single host [25]. The resulting intra-host diversity has been shown to affect disease progression [26], tissue tropism [27], transmission risk [28], transmission heterogeneity [29], and treatment outcome [30,31] in various RNA viruses. Recent findings indicate that the mutation rate of SARS-CoV-2 is about as high as the one observed in the SARS-CoV genome ($0.80–2.38 \times 10^{-3}$ nucleotide substitutions per site per year) [32,33]. Although coronaviruses have evolved a proofreading capability attributed to Nonstructural protein 14 resulting in lower mutation rates than other positive-sense ssRNA viruses [34–36], the study of the intra-host genetic diversity of the novel SARS-CoV-2 virus remains important to gain a deeper understanding of its evolution and transmission dynamics and possible implications on its pathology. Previous studies highlight that intra-host genetic diversity in clinical samples is indeed prevalent, and some genomic regions that are susceptible to alterations in the SARS-CoV-2 virus have been identified [37,38]. Accounting for intra-host diversity can also improve the resolution of phylogenomic analyses of SARS-CoV-2 [39] and our ability to detect selection pressure [40].

To assess the within-host genetic diversity of different viral variants, deep-coverage sequencing is needed to access low-frequency subclonal mutations. However, calling low-frequency mutations reliably from deep sequencing data remains challenging because of various amplification and sequencing errors. Several computational methods have been proposed for this task [41,42], and specialised bioinformatics pipelines have been developed to streamline and automate the analysis from raw sequencing reads to consensus sequences and single-nucleotide variants [43,44]. Here, we employ V-pipe [44], a workflow that includes quality control, read mapping and mutation calling, to robustly identify genetic diversity of viral genomes. We examine the intra-host diversity across a large cohort of 3939 public SARS-CoV-2 samples and, as an independent validation cohort, 749 additional samples which have recently been sequenced to study the spread of SARS-CoV-2 in Switzerland [45]. We analyze the distribution of genetic diversity across the genome, including uncovering highly diverse bases and small regions, and we explore the diversity across patient samples. By integrating technical and epidemiological covariates, we show that within-patient viral genetic diversity increases with age of the infected patient.
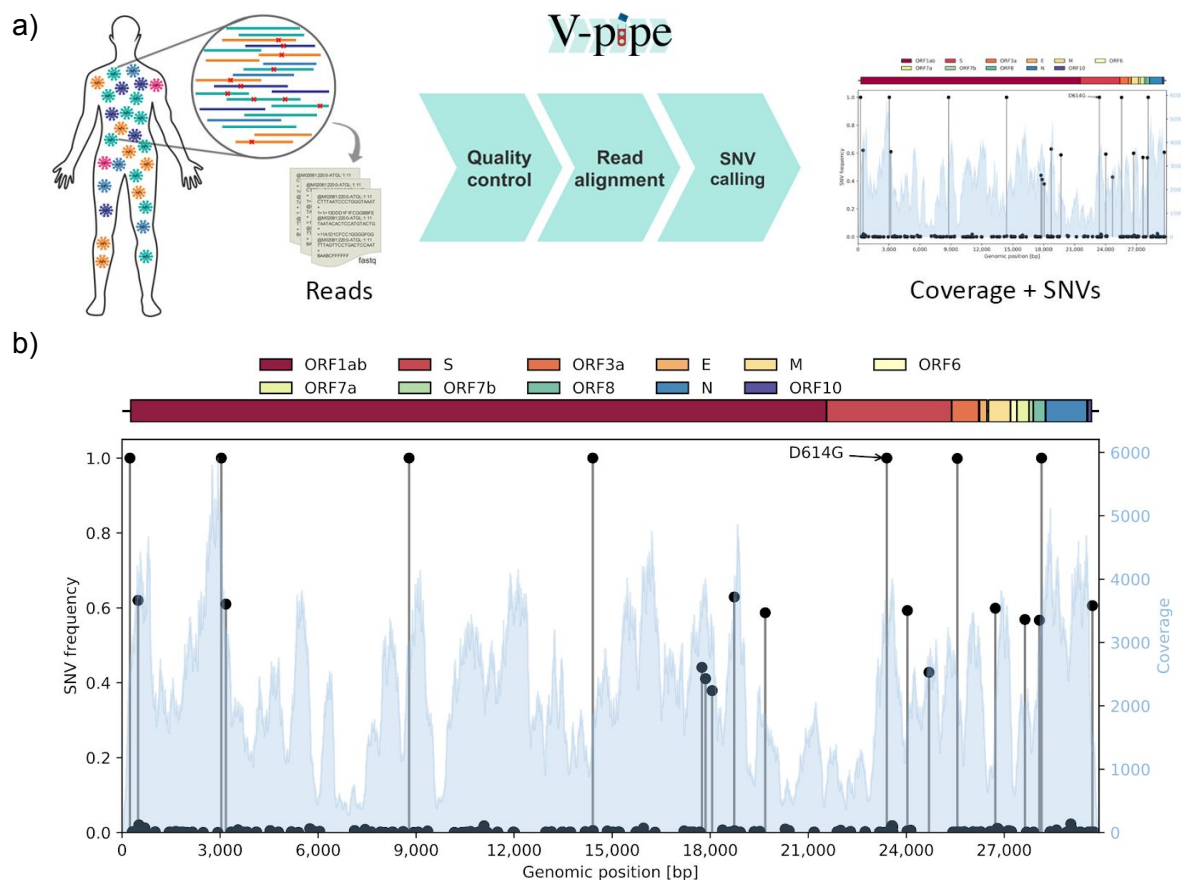
Figure 1: (a) Different viral variants may coexist in the same host such that the sequencing reads contain a mixture of the different components and their SNVs. We employ the bioinformatics pipeline V-pipe for automated end-to-end processing of the raw sequencing data and calling of SNVs and their frequencies [44]. (b) An example output of the workflow for a single sample (SRR11953858) displaying variations in the coverage (blue histogram, right y-axis) and the frequencies of the different SNVs (black lollipops, left y-axis) across the genome (x-axis). Along with clonal mutations with frequencies near 100%, and many low-frequency variants, several other mutations have frequencies of around 40% and 60% in this example.

Results

We downloaded sequencing data from the SRA from a total of 5934 SARS-CoV-2 samples, which had undergone deep sequencing with Illumina technology. Samples with low (median below 1000) or highly variable (lower quartile below 100 or upper quartile above 10,000) coverage were removed (Figure S1). The remaining 3940 samples were processed using V-pipe [44] (Figure 1a). For each sample, we obtained the coverage across the genome and all mutated positions, including all variant bases and their frequencies (Figure 1b). A single sample with no detected mutations at all was removed to leave a final public cohort of 3939 SARS-CoV-2 samples. In addition, we analyze a new set of 749 sequences derived from samples collected in Switzerland. These sequences were generated as part of our sequencing effort described in [45]. While [45] uses only the consensus sequences for a phylodynamic analysis, here we directly analyse the entire set of raw deep-sequencing reads to assess intra-host diversity. This uniformly produced set of sequences, which we make public on the SRA, serves as a validation cohort.
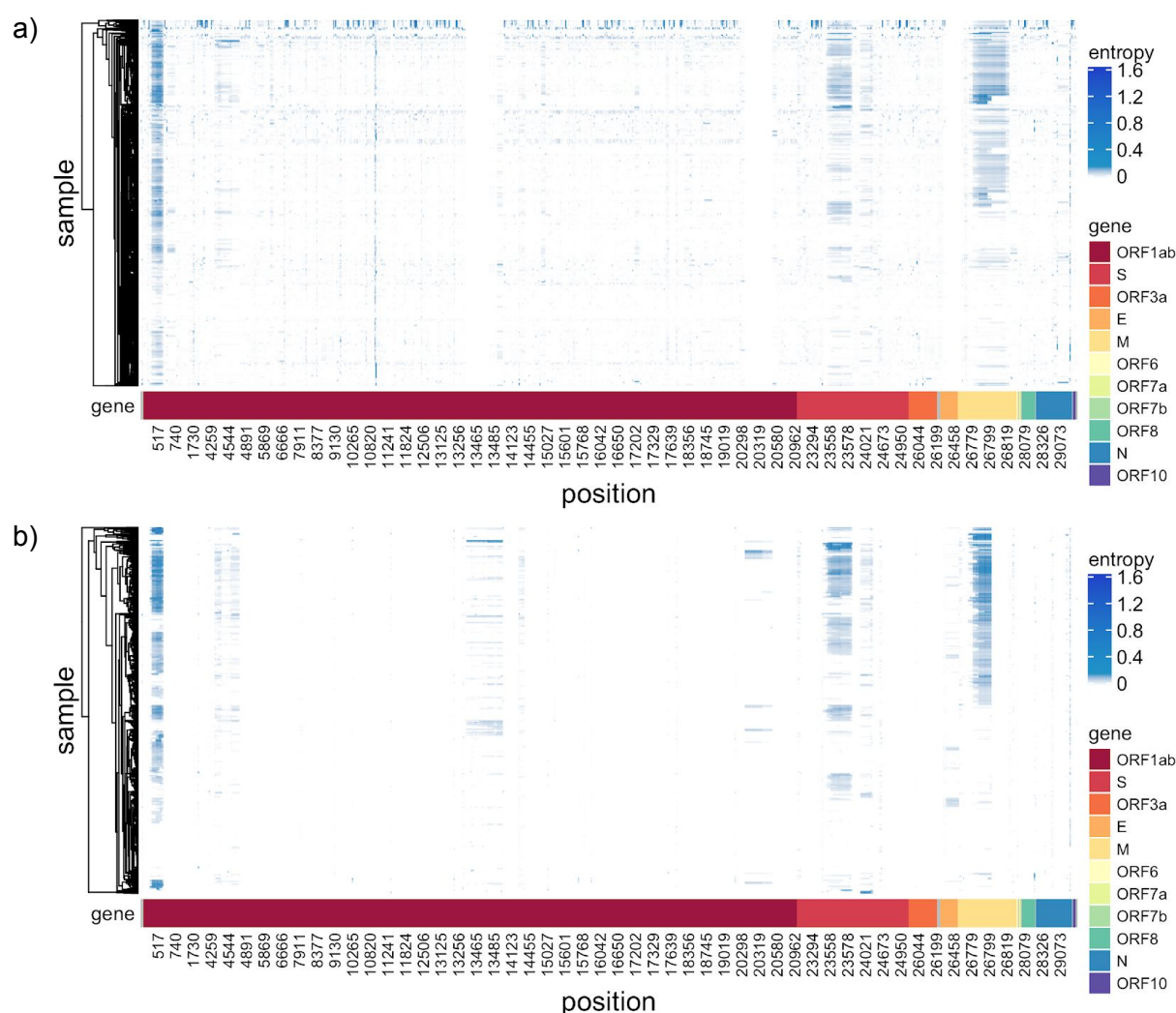
Figure 2: Intra-host diversity measured as entropy of the nucleotide distribution across genome and samples. Only positions with diversity in at least 4% of the samples are selected for each cohort, with the union of positions displayed for comparison. (a) the public cohort, (b) the data from Switzerland.

As a measure of intra-host genetic diversity, we computed the entropy of the nucleotide distribution for each sample and each genomic locus (Materials and Methods). Low entropy indicates a highly conserved site in the patient's virus population, while high entropy is indicative of variation among nucleotides.

The quartiles of the coverage over the public samples (Figure S2) display typical coverages of over a thousand reads at each genomic site (in line with the coverage filters) meaning that we can detect low-frequency SNVs down to 1-2% [46] (or using estimates from the Lander-Waterman model [47]). However, there are regular dips at the primer locations [48], and noticeable wider dips between genomic positions around 19 to 23 kb. This region covers the end of gene *ORF1ab* encoding the endoribonuclease and 2'-O-methyltransferase, as well as much of the *S* gene producing the S1 subunit, needed for the initial attachment of the virus to the target cell.

We observe a wide spectrum of diversity both across the genome and across samples (Figure 2, focussing on the positions across the genome mutated in at least 4% of the samples). In particular there are subclonal regions visible (as darker contiguous bars in Figure 2) at the start of *ORF1ab*, in the Spike *S* gene and across much of the Matrix *M* gene. There are also highly diverse individual bases and small regions (Tables S1-S4), particularly in the public data.

Next, we summarize genetic diversity per gene by computing the average entropy across all positions of the gene. We observe roughly similar diversity for each gene in the public cohort, except for gene *M* which is much more diverse, and *ORF7b* displays hardly any diversity (Figure 3a), while we observe the same trends, but larger differences across genes in the Swiss cohort (Figure 3b).

The most diverse site in the public data (Table S1 and Figure 3a), which has clonal or subclonal mutations in nearly half the samples, is at position 11075 in the region of the *ORF1ab* gene coding for non-structural protein 6, a transmembrane protein containing 7 transmembrane helices. Most of the diversity at this position is derived from low-frequency deletions, which would alter downstream amino acids and introduce a premature stop codon. Among the mutated samples, deletions occur at an average rate of 2.62%. 12.5% of mutated samples include variants with a T>C substitution, which occurs with an average frequency of 1.75% among those samples (corresponding to an overall average rate of 0.22% among all mutated samples). In the reference genome, the codon incorporates a Phenylalanine amino acid (Phe35 of nsp6), however, any variation here will lead to a change in the inserted amino acid; for example, a T>C mutation results in the incorporation of a Leucine.

Another residue in close proximity to Phe35, namely Leu37 of nsp6, has often been found to be replaced by a Phenylalanine residue in recent sequences from Europe, Asia and America [49]. Here, we also identify a high entropy at position 11083 inside the codon of Leu37 (Table S1), which is affected in around a quarter of samples. Any alteration from the reference base G to either a T or C results in a Phenylalanine residue as opposed to a Leucine. Here we observe a large average frequency of Thymine bases (44.65%) and common deletions (4.93%) which also lead to a Phenylalanine residue and a premature stop codon.

Genomic position 23403 is the site at which an A>G mutation causes the well-known D614G amino acid alteration in the Spike region. Studies suggest that this alteration provides a fitness advantage and becomes the more prevalent variant in the population over time [50], and it is also the dominant variant in our public cohort. We find that it is among the 100 most diverse positions in the SARS-CoV-2 genome (rank 95 in our cohort) with 36.0% of our samples having mutations there relative to our cohort consensus G, and 29.7% having diversity of different variants coexisting. Among the mutations, there is an average frequency of 90.4% A, 9.5% G and 0.1% deletions. In the Swiss cohort, position 23403 has the second highest diversity among all individual bases (Table S2).
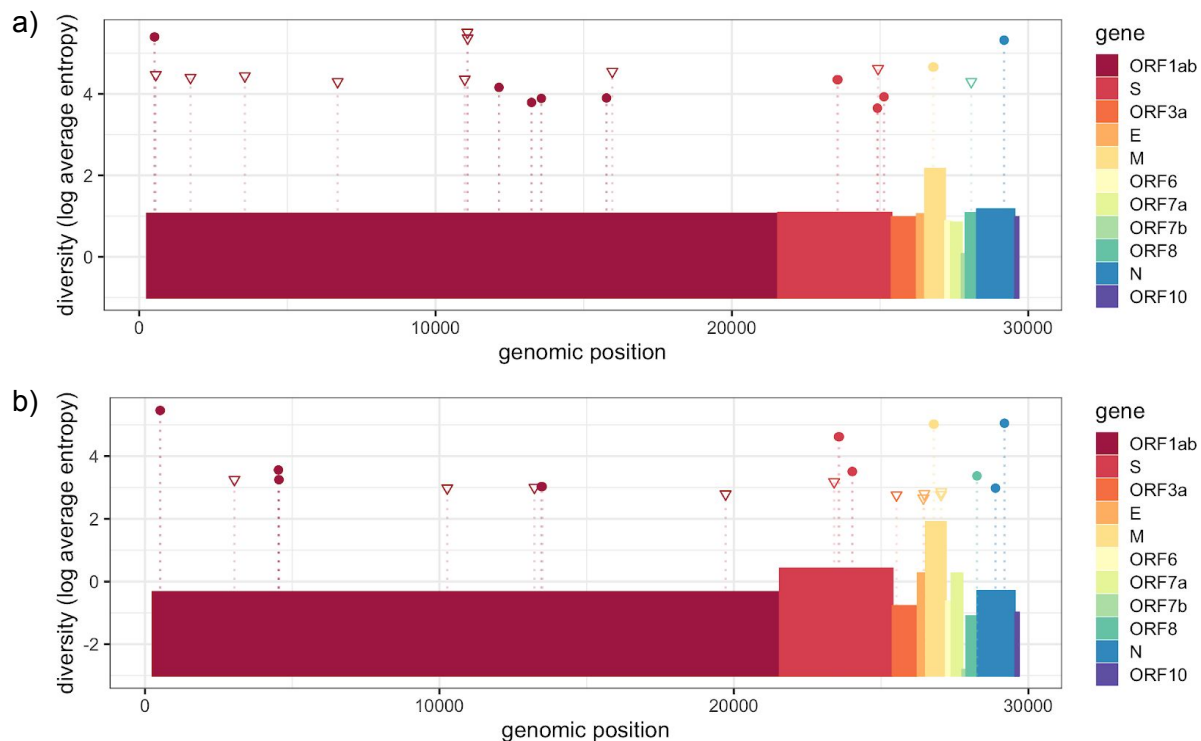
Figure 3: The average entropy per gene (boxes), log transformed, along with the 10 most diverse (in terms of cumulative entropy over samples) positions (triangles) and the 10 most diverse consecutive regions (dots) from Tables S1 and S3 for the public data in (a) and Tables S2 and S4 for the data from Switzerland in (b).

We also assessed genetic diversity over consecutive genomic regions (Materials and Methods). The most diverse regions in both the public and Swiss data (Tables S3 and S4, adjacent sites all with a log entropy above 3.0 for the public data and above 2.0 for the Swiss data) include a span of 16 bases at the start of the *ORF1ab* gene (Figure 3). This region forms part of the nsp1 protein which shuts down host mRNA translation and protein production via interactions with the 40S and 80S ribosomal subunits. A notable consequence of this nsp1 induced inhibition of translation and protein production is the reduction in innate immune response activity, specifically interferon signalling [51].

The next most diverse genomic region (Tables S3 and S4) consists of two adjacent positions at 29187 and 29188 within a single codon in the *N* gene forming the nucleoprotein. The reference bases at those two positions are C and A, respectively, which results in the incorporation of an Alanine residue at position 305 of the nucleoprotein (Ala305). However, due to the redundancy of the genetic code, alterations in position 29188 alone do not lead to an alteration on the protein level.

The remaining regions in Table S3 with a length greater than two lie within the *M* and *S* genes, and are similarly present in the Swiss cohort (Table S4). The 41 nucleotides from 26780 to 26820 lead to the incorporation of 15 amino acids from Cys86 to Phe100 forming a section of the matrix protein's transmembrane region. The matrix protein is known to be central in viral assembly with several interacting partners including itself, the envelope protein, the nucleoprotein and the spike protein [52,53]. The spike protein, when primed, is divided into the S1 and S2 subunits with S1 being essential for hACE2 recognition and S2

mediating viral entry into the host cell. Another region of high entropy comprises 30 nucleotides within the Spike gene which incorporates 11 amino acids to the S1 subunit of SARS-Cov-2 ranging from Ile664 to Tyr674. These amino acids lie between the receptor binding domain and Furin cleavage site needed for S protein priming [54,55].

The distribution of total entropy per base has a long tail (Figure 4a) resulting in locations with extremely high diversity, such as the examples discussed above, while the majority of the genome is relatively conserved. Taking a logarithmic transform of the entropy we can observe much more detail in the distribution, with a roughly normal distribution and a slight negative skew in the public data (Figure 4a inset). Considering the entropy per sample instead provides a picture of how diversity is distributed among samples. We again find a long tail (Figure 4b) with certain individuals having vastly more internal diversity than the majority of samples dominated by clonal variants and low-frequency mutations. Detecting diversity in each sample is heavily dependent on the sequencing technology, sample processing, and sequencing depth. For example, all the most diverse samples (Table S5) come from one SRA study (SRP253798) which makes up 29% of the entire cohort. The samples with the highest detected diversity had between 10% and 30% of the genome affected, with the vast majority of their mutations (over 99.8%) being subclonal and the coexistence of more than one character (base or deletion) in the mapped reads.

Like the entropy, the number of clonally or subclonally mutated positions per sample also has a long tail (Figure S3) with, in the public data, for example, quartiles at 119, 211 and 378.5 positions being mutated (corresponding to 0.40%, 0.71% and 1.27% of the genome) but a maximum of 8705 positions (29.11% of the genome).

Finally, we tested the hypothesis that epidemiological parameters are related to viral genetic diversity. Specifically, to determine whether the host's age, sex or geographical location predict the diversity of their virus population, we performed regression modelling on the subset of samples for which we have such information. This resulted in 1043 samples from Australia which were sequenced with paired-end amplicon sequencing with PCR amplification. We also adjusted for technical parameters of the sequencing to avoid confounding (Materials and Methods). We found that sex is not a significant predictor of diversity (p = 0.50). By contrast, age is significantly associated with intra-host viral genetic diversity (p = 4 × 10^{-4}) with each decade increasing the total entropy by 8.6% on average (Table 1, Figure S5).

Clinical covariates were not available for the Swiss cohort, but sequencing date was utilized as a proxy for decreasing age, since testing expanded to younger populations over time (Figure S6). The significant decrease in diversity over time (p < 10^{-7}; Table 2, Figure S7) in the Swiss cohort therefore corroborates the association of increased diversity with age uncovered in the public cohort.
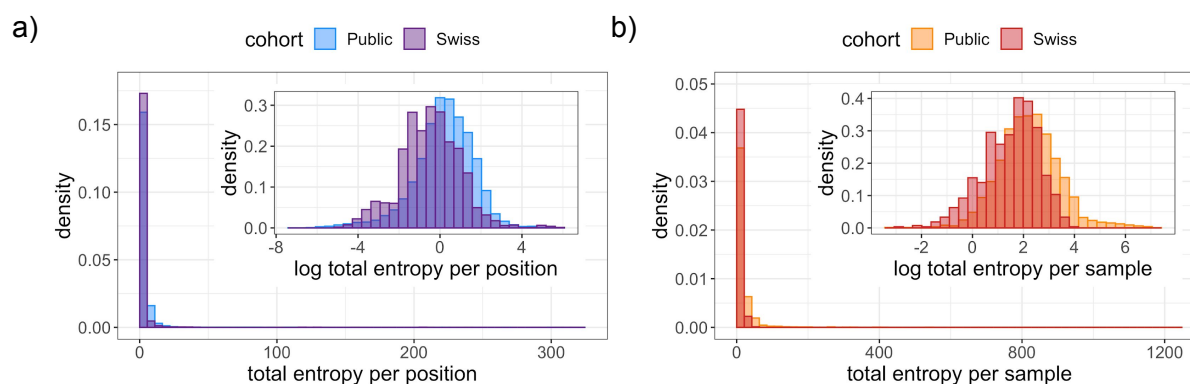
Figure 4: The distributions of the total entropy in the public and Swiss cohorts per position (a) and per sample (b). Insets: Under a logarithmic transformation, the distribution per position (a) and per sample (b).

Discussion

We processed a large cohort of 3939 deeply sequenced public SARS-CoV-2 genomes, and 749 samples from Switzerland, with the bioinformatics software V-pipe [44] to uncover within-patient genetic diversity. We observe a heavy tail distribution in diversity per sample and per position indicating that much of the diversity is concentrated in small numbers of sites and patients. The most diverse small regions were consistent between the public and Swiss data, though the individual bases were more varied.

Detecting and quantifying intra-patient genetic diversity from deep sequencing data is technically challenging. It may be heavily influenced by the sample preparation and how it is sequenced, along with possible artifacts arising from the process, so that extremely diverse patients may not be comparable across cohorts. Accounting for such technical parameters of the sequencing process and coverage, which affects the detection limit of diversity, we find that age is a significant predictor of diversity in the public data. The model predicts that on average genetic diversity increases by 8.6% every ten years. The increase in diversity with host age is corroborated in the Swiss data. Age has previously been associated strongly with worse disease outcome and higher death rates [56], along with concomitant comorbidities [57]. With high-quality clinical and genetic data, it will be highly relevant to see whether diversity is a cause or a consequence of disease progression. Likewise, with future transmission network data it will be interesting to uncover whether diversity increases infectiousness, as for influenza [28].

The detection of subclonal mutations is affected by the sequencing depth at each position, and across the genome depending on the amplification and capture of RNA for sequencing some regions may be more poorly resolved. The coverage distribution therefore affects our ability to detect highly diverse bases. With this caveat, the most diverse gene is the Matrix *M* gene while highly diverse positions include a mix of low-frequency variants common to a quarter of the cohort or more, and rarer high-frequency subclonal mutations in around 5% of the cohort. The observation of common low-frequency and less common high-frequency genetic variants is in line with previous research on both intra- and inter-host genetic diversity of SARS-CoV-2 [21,39]. The D614G variant, which appears to increase infectivity and is becoming more dominant over time [50] is the dominant variant in our public

cohort. It also exhibits high intra-host diversity with 29.7% of the cohort experiencing subclonal mutations with the different variants coexisting. This diversity is mimicked in the data from Switzerland, where the D614G variant is actually encoded by the second most diverse genomic position.

Materials and Methods

Public data

We retrieved data from the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) on June 10, 2020. Samples with the term '"Severe acute respiratory syndrome coronavirus 2"[Organism] OR "Sars-Cov-2[All Fields]"' were filtered and only one copy of duplicates with the same BioSample ID was retained. We used the meta file to further filter the samples by Illumina technology. This resulted in 5934 samples which were associated with downloadable data.

Subsequent to downloading the selected sample set, we trimmed all read files using PRINSEQ ([58] version 0.20.4, parameters: -ns_max_n 4 -min_qual_mean 30 -trim_qual_left 30 -trim_qual_right 30 -trim_qual_window 10 -min_len <80% of average read length>), mapped them to NC_045512.2 using bwa ([59] version 0.7.17-r1188, subcommand: mem). Coverage quartiles for each sample are displayed in Figure S1.

Data from samples collected in Switzerland

Sample collection and sequencing are detailed in [45]. Briefly, we obtained RNA samples extracted from nasal swab tests which had previously tested positive on RT-qPCR from Viollier AG laboratory and sequenced them at the Genomics Facility Basel. We performed reverse transcription using random hexamers and PCR amplified the resulting DNA with primers from the artic-ncov2019 protocol [https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3]. We prepared libraries from these 4000 bp-long amplicons using Illumina TruSeq adapter sequences and sequenced them on an Illumina MiSeq System (Paired-end sequencing, 2x 251 cycle). The phylogenetic relationship between the consensus sequence of 681 samples from this collection has been previously analyzed as part of the subset of GISAID data available for Switzerland until July 10, 2020 in [24] and results based on the consensus sequences of the full dataset are presented in [45]. Here we focus on the raw reads directly and analyse the deep sequencing data to uncover within host diversity. The collection of swabs analysed here spans a time period from Mar 4, 2020 to August 13, 2020, with samples from across Switzerland. The analysed reads are available in the SRA (see Data Availability below).

Data on the age distribution of COVID-19 cases in Switzerland was downloaded from [https://www.bag.admin.ch/dam/bag/de/dokumente/mt/k-und-i/aktuelle-ausbrueche-pandemien/2019-nCoV/covid-19-basisdaten-fallzahlen.xlsx.download.xlsx/Dashboards_1&2_COVID19_swiss_data_pv.xlsx] on September 21, 2020.

Filtering

Samples were subset by applying a coverage filter (minimum lower quartile: 100, minimum median: 1000, maximum upper quartile: 10000). After this filtering, 3940 public samples and 749 samples from Switzerland were retained for later analyses.

Data processing

We used V-pipe ([44]; sars-cov2 branch of https://github.com/cbg-ethz/V-pipe) to call variants for each sample using ShoRAH [60] and default settings, including discarding deletions with a frequency below 0.5%. One sample had no remaining variants detected and was excluded giving a final cohort size of 3939. For each called variant per sample, we computed the relative frequency f_k of each character k (nucleotide A, C, G, T, or deletion) among the mapped reads at that position. The entropy is then computed as

$$H = -\sum_{k=1}^{5} f_k \log(f_k)$$

summarising the five frequencies in a single measure of diversity. The entropy is zero whenever one character has a frequency of 100%, and is maximised when all the characters have equal frequency 1/5, giving a maximum value of log(5) ≈ 1.61. We denote the entropy of sample i at base j by H_{ij} and compute the total entropy per position in the genome by summing over samples $H_j = \sum_i H_{ij}$ (Table S1 and S2). For the data from Switzerland we additionally multiply by the ratio of cohort sizes (3939/749) to make the total entropy values per position comparable across the cohorts. We compute the average entropy over a consecutive genomic region J as $\frac{1}{|J|} \sum_{j \in J} H_j$, and take the logarithm to obtain the log average entropy (Table S3 and S4). We compute the total entropy per sample by summing over bases $H_i = \sum_j H_{ij}$ (Table S5).

Regression modelling

To evaluate which covariates are predictive of diversity, we build a regression model on the public data of log total entropy on age, sex, and country of origin. To adjust for the possible effects of coverage and sequencing technology on the diversity, we include factors for paired-end or single-end sequencing, the assay type, the library selection and the SRA study. We also include the logarithm of the median coverage. Since not only the average coverage, but its variability may affect the ability to detect SNVs, we further include the IQR of the coverage, divided by the median m, and again log transformed:

$$\log(\text{entropy}) \sim \text{age} + \text{sex} + \text{country} + \text{sequencing factors} + \log(\text{median coverage}) + \log(\text{IQR}/m)$$

Filtering the samples which have age and sex information left 1060 samples, of which all but 17 were from Australia. We therefore retained just those 1043 from Australia, and removed the country dependence from the regression. All remaining samples were paired-end amplicon sequencing with PCR amplification from the study SRP253798, so those factors were also removed to provide the final regression:

$$\log(\text{entropy}) \sim \text{age} + \text{sex} + \log(\text{median coverage}) + \log(\text{IQR}/m)$$

Of the 1043 samples, 480 (46.0%) were female and 563 (54.0%) were male, while the distribution of ages (Figure S4) has a median value of 46 and lower and upper quartiles at 29 and 60. The results of the regression are listed in Table 1.

Predictor	Coefficient	Standard error	t-statistic	p-value
sex (male)	-0.0563	0.0839	-0.671	0.50
age (decade)	0.0822	0.0231	3.56	0.00038
log(median coverage)	0.190	0.125	1.51	0.13
log(IQR/m)	1.48	0.122	12.1	$< 2 \cdot 10^{-16}$

Table 1: regression coefficients of log entropy against the above predictors for the 1043 public samples with covariate information.

For the data from Switzerland, we do not have clinical covariates, only the date of sequencing. However, we can use the sequencing date as a proxy for age, because over time, the age distribution has progressively reduced in Switzerland as a whole (Figure S6). All samples were processed in the same way, so for the regression we model

$$\log(\text{entropy}) \sim \text{date} + \log(\text{median coverage}) + \log(\text{IQR}/m)$$

with the results in Table 2.

Predictor	Coefficient	Standard error	t-statistic	p-value
date (30 days)	-0.108	0.019	-5.68	$1.95 \cdot 10^{-8}$
log(median coverage)	0.196	0.107	1.83	0.068
log(IQR/m)	1.41	0.131	10.8	$< 2 \cdot 10^{-16}$

Table 2: regression coefficients of log entropy against the above predictors for the 749 Swiss samples.

For visualisation purposes, we regress patient age on sample date from data collected about positive tests in Switzerland as a whole (Figure S6). The model is then used to predict the age of our 749 deeply sequenced samples based on the sequencing date. Against the predicted age, we plot the log total entropy, adjusted for the coverage covariates, to show how the negative correlation with date (Table 2) corroborates an increase in entropy with age (Figure S7).

Code availability

The source code to process the samples and perform and reproduce the analyses is available on GitHub (https://github.com/cbg-ethz/SARS-CoV-2_Analysis) in the form of multiple Snakemake [61] workflows.

Data availability

The public data is available from the SRA, as described in the Methods and Materials section. The Swiss data has been added to the SRA with accession number PRJEB38472, scheduled to be publicly available from November 1, 2020.

Author contributions

Bibliography

1. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020;94. doi:10.1128/JVI.00127-20

2. Ong SWX, Tan YK, Chia PY, Lee TH, Ng OT, Wong MSY, et al. Air, Surface Environmental, and Personal Protective Equipment Contamination by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) From a Symptomatic Patient. JAMA. 2020; doi:10.1001/jama.2020.3227

3. de Lusignan S, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, et al. Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. Lancet Infect Dis. 2020; doi:10.1016/S1473-3099(20)30371-6

4. Folegatti PM, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, et al. Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. Lancet. 2020;396: 467–478. doi:10.1016/S0140-6736(20)31604-4

5. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med. 2020;26: 1037–1040. doi:10.1038/s41591-020-0916-2

6. Long Q-X, Tang X-J, Shi Q-L, Li Q, Deng H-J, Yuan J, et al. Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. Nat Med. 2020;26: 1200–1204. doi:10.1038/s41591-020-0965-6

7. Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-M, et al. Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. J Clin Med. 2020;9. doi:10.3390/jcm9020538

8. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020;172: 577–582. doi:10.7326/M20-0504

9. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;20: 533–534. doi:10.1016/S1473-3099(20)30120-1

10. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5: 536–544.

doi:10.1038/s41564-020-0695-z

11. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269. doi:10.1038/s41586-020-2008-3

12. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nat Microbiol. 2020;5: 562–569. doi:10.1038/s41564-020-0688-y

13. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273. doi:10.1038/s41586-020-2012-7

14. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020;181: 271-280.e8. doi:10.1016/j.cell.2020.02.052

15. Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, et al. SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. Cell. 2020;181: 1016-1035.e19. doi:10.1016/j.cell.2020.04.035

16. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26: 450–452. doi:10.1038/s41591-020-0820-9

17. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020;583: 286–289. doi:10.1038/s41586-020-2313-x

18. Xiao M, Liu X, Ji J, Li M, Li J, Yang L, et al. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. Genome Med. 2020;12: 57. doi:10.1186/s13073-020-00751-4

19. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22: 30494. doi:10.2807/1560-7917.ES.2017.22.13.30494

20. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34: 4121–4123. doi:10.1093/bioinformatics/bty407

21. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med. 2020;18: 179. doi:10.1186/s12967-020-02344-6

22. Jia Y, Shen G, Zhang Y, Huang K-S, Ho H-Y, Hor W-S, et al. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. BioRxiv. 2020; doi:10.1101/2020.04.09.034942

23. Phan T. Genetic diversity and evolution of SARS-CoV-2. Infect Genet Evol. 2020;81: 104260. doi:10.1016/j.meegid.2020.104260

24. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European

Region, January to June 2020. Euro Surveill. 2020;25. doi:10.2807/1560-7917.ES.2020.25.32.2001410

25. Lauring AS, Andino R. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. 2010;6: e1001005. doi:10.1371/journal.ppat.1001005

26. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature. 2006;439: 344–348. doi:10.1038/nature04388

27. Tsibris AMN, Korber B, Arnaout R, Russ C, Lo C-C, Leitner T, et al. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. PLoS ONE. 2009;4: e5683. doi:10.1371/journal.pone.0005683

28. Poon LLM, Song T, Rosenfeld R, Lin X, Rogers MB, Zhou B, et al. Quantifying influenza virus diversity and transmission in humans. Nat Genet. 2016;48: 195–200. doi:10.1038/ng.3479

29. Zhang Y, Leitner T, Albert J, Britton T. Inferring transmission heterogeneity using virus genealogies: Estimation and targeted prevention. PLoS Comput Biol. 2020;16: e1008122. doi:10.1371/journal.pcbi.1008122

30. Hu Y, Lu S, Song Z, Wang W, Hao P, Li J, et al. Association between adverse clinical outcome in human disease caused by novel influenza A H7N9 virus and sustained viral shedding and emergence of antiviral resistance. Lancet. 2013;381: 2273–2279. doi:10.1016/S0140-6736(13)61125-3

31. Wensing AM, Calvez V, Ceccherini-Silberstein F, Charpentier C, Günthard HF, Paredes R, et al. 2019 update of the drug resistance mutations in HIV-1. Top Antivir Med. 2019;27: 111–121.

32. Sapoval N, Mahmoud M, Jochum MD, Liu Y, Leo Elworth RA, Wang Q, et al. Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission. BioRxiv. 2020; doi:10.1101/2020.07.02.184481

33. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol. 2004;4: 21. doi:10.1186/1471-2148-4-21

34. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells. 2020;9. doi:10.3390/cells9051267

35. Peck KM, Lauring AS. Complexities of viral mutation rates. J Virol. 2018;92. doi:10.1128/JVI.01031-17

36. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol. 2011;8: 270–279. doi:10.4161/rna.8.2.15013

37. Rose R, Nolan DJ, Moot S, Feehan A, Cross S, Garcia-Diaz J, et al. Intra-host site-specific polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. medRxiv. 2020; doi:10.1101/2020.04.24.20078691

38.   Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, Mentis A. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. BioRxiv. 2020; doi:10.1101/2020.03.27.009480

39.   Ramazzotti D, Angaroni F, Maspero D, Gambacorti-Passerini C, Antoniotti M, Graudenzi A, et al. Quantification of Intra-Host Genomic Diversity of SARS-CoV-2 Allows a High-Resolution Characterization of Viral Evolution and Reveals Functionally Convergent Variants. BioRxiv. 2020; doi:10.1101/2020.04.22.044404

40.   Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. BioRxiv. 2020; doi:10.1101/2020.07.06.189944

41.   Posada-Cespedes S, Seifert D, Beerenwinkel N. Recent advances in inferring viral diversity from high-throughput sequencing data. Virus Res. 2017;239: 17–32. doi:10.1016/j.virusres.2016.09.016

42.   Knyazev S, Hughes L, Skums P, Zelikovsky A. Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. Brief Bioinformatics. 2020; doi:10.1093/bib/bbaa101

43.   Lee ER, Parkin N, Jennings C, Brumme CJ, Enns E, Casadellà M, et al. Performance comparison of next generation sequencing analysis pipelines for HIV-1 drug resistance testing. Sci Rep. 2020;10: 1634. doi:10.1038/s41598-020-58544-z

44.   Posada-Céspedes S, Seifert D, Topolsky I, Metzner KJ, Beerenwinkel N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput sequencing data. BioRxiv. 2020; doi:10.1101/2020.06.09.142919

45.   Nadeau S, Beckmann C, Topolsky I, Vaughan T, Hodcroft E, Redondo M, et al. Quantifying SARS-CoV-2 spread in Switzerland based on genomic sequencing data . In preparation. 2020;

46.   McElroy K, Zagordi O, Bull R, Luciani F, Beerenwinkel N. Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias. BMC Genomics. 2013;14: 501. doi:10.1186/1471-2164-14-501

47.   Eriksson N, Pachter L, Mitsuya Y, Rhee S-Y, Wang C, Gharizadeh B, et al. Viral population estimation using pyrosequencing. PLoS Comput Biol. 2008;4: e1000074. doi:10.1371/journal.pcbi.1000074

48.   Baker DJ, Kay GL, Aydin A, Le-Viet T, Rudder S, Tedim AP, et al. CoronaHiT: large scale multiplexing of SARS-CoV-2 genomes using Nanopore sequencing. BioRxiv. 2020; doi:10.1101/2020.06.24.162156

49.   Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, Cauda R, et al. Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. J Infect. 2020;81: e24–e27. doi:10.1016/j.jinf.2020.03.058

50.   Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020; doi:10.1016/j.cell.2020.06.043

51. Thoms M, Buschauer R, Ameismeier M, Koepke L, Denk T, Hirschenberger M, et al. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. Science. 2020; doi:10.1126/science.abc8665

52. Arndt AL, Larson BJ, Hogue BG. A conserved domain in the coronavirus membrane protein tail is important for virus assembly. J Virol. 2010;84: 11418–11428. doi:10.1128/JVI.01131-10

53. J Alsaadi EA, Jones IM. Membrane binding proteins of coronaviruses. Future Virol. 2019;14: 275–286. doi:10.2217/fvl-2018-0144

54. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res. 2020;176: 104742. doi:10.1016/j.antiviral.2020.104742

55. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nat Commun. 2020;11: 1620. doi:10.1038/s41467-020-15562-9

56. Onder G, Rezza G, Brusaferro S. Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy. JAMA. 2020; doi:10.1001/jama.2020.4683

57. Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int J Infect Dis. 2020;94: 91–95. doi:10.1016/j.ijid.2020.03.017

58. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27: 863–864. doi:10.1093/bioinformatics/btr026

59. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. 2013 [cited 24 Jul 2020]. Available: https://arxiv.org/abs/1303.3997

60. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 2011;12: 119. doi:10.1186/1471-2105-12-119

61. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522. doi:10.1093/bioinformatics/bts480

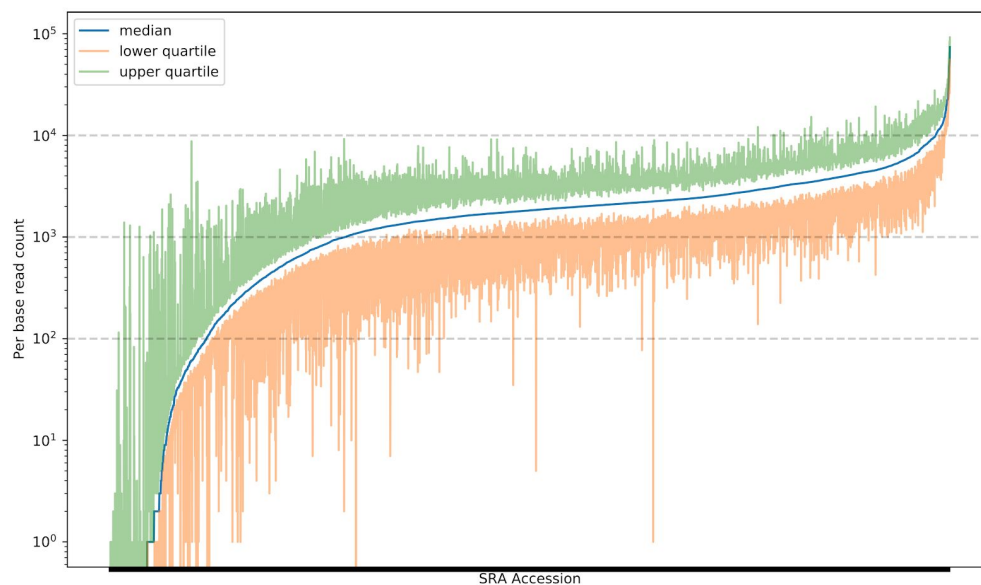# Supplementary Figures and Tables

Figure S1. The median coverage, along with the lower and upper quartiles for the 5934 samples downloaded from the Sequence Read Archive. The dashed grey lines correspond to the coverage filters used to subset the samples before further processing (minimum lower quartile: 100, minimum median: 1000, maximum upper quartile: 10000).
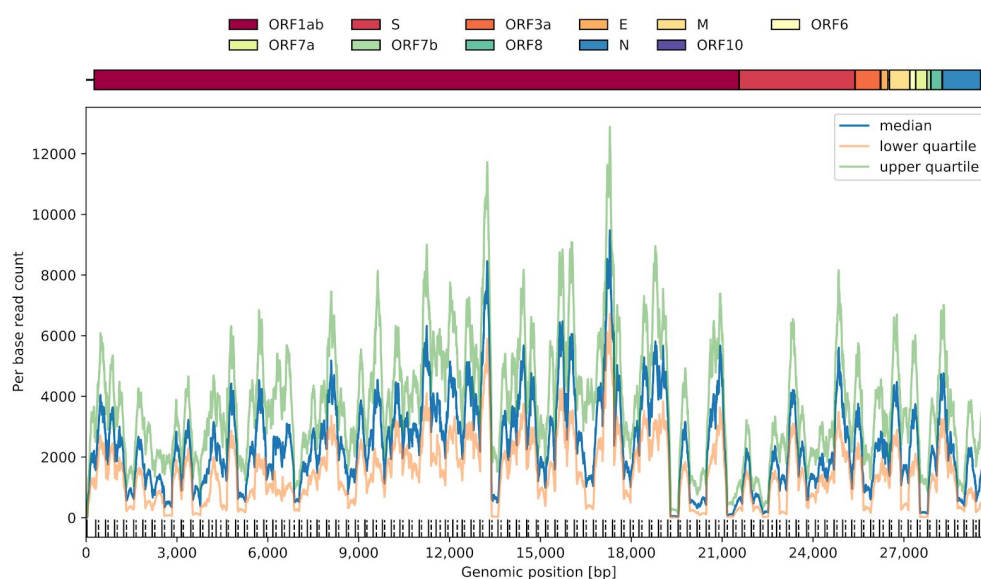

Figure S2: The lower and upper quartiles of the coverage per genomic position, along with the median coverage, among the 3940 coverage-filtered public sample sequences. The locations of typical primers are indicated along the bottom row (solid at the start of left primers and dashed at the end of right primers).

position	gene	log entropy	mutated samples (%)	ref. base	A (%)	C (%)	G (%)	T (%)	del. (%)
11075	ORF1ab	5.51	49.30	T	0.00	0.22	0.00	97.16	2.62
11083	ORF1ab	5.37	24.37	G	0.01	0.01	50.40	44.65	4.93
24933	S	4.62	6.93	G	0.02	0.92	78.08	20.54	0.44
15965	ORF1ab	4.55	28.21	G	0.00	0.00	98.31	1.35	0.34
558	ORF1ab	4.47	5.97	G	0.04	0.41	82.82	16.37	0.36
3564	ORF1ab	4.44	4.32	G	0.08	0.34	64.54	34.63	0.42
1730	ORF1ab	4.40	24.45	G	1.84	0.00	98.16	0.00	0.00
10986	ORF1ab	4.36	5.69	G	0.91	0.27	88.20	9.19	1.43
6696	ORF1ab	4.30	22.34	C	0.01	96.92	0.00	2.97	0.10
28079	ORF8	4.30	4.72	G	0.00	0.13	83.48	14.62	1.76

Table S1: The 10 most diverse positions in the genome in the public data ranked by their entropy, along with the fraction of samples exhibiting any mutation, and, for those samples, the distribution of average mutation frequencies across the different bases or deletion.

position	gene	log entropy	mutated samples (%)	ref. base	A (%)	C (%)	G (%)	T (%)	del. (%)
3037	ORF1ab	3.25	21.50	C	0	5.25	0	94.61	0.14
23403	S	3.18	43.39	A	3.47	0	96.46	0.04	0.02
13225	ORF1ab	3.00	7.61	C	0	94.10	0	5.89	0.01
10265	ORF1ab	2.98	10.68	G	91.02	0	8.83	0	0.14
27040	M	2.86	25.77	C	0.25	99.75	0	0	0
26465	E	2.8	8.81	T	0	0	0	99.17	0.83
19718	ORF1ab	2.79	10.01	C	0	2.49	0	97.49	0.02
27033	M	2.77	24.97	G	0	0	99.76	0.24	0
25521	ORF3a	2.76	1.74	C	0	85.74	0	14.26	0
26434	E	2.66	21.9	A	99.76	0	0.24	0	0

Table S2: The 10 most diverse positions in the genome in the Swiss data ranked by their entropy, along mutation patterns for those positions across the cohort.

start	end	length	gene	log average entropy
508	523	16	ORF1ab	5.4
29187	29188	2	N	5.32
26780	26820	41	M	4.66
23553	23582	30	S	4.35
12140	12141	2	ORF1ab	4.16
25134	25135	2	S	3.93
15768	15769	2	ORF1ab	3.9
13570	13571	2	ORF1ab	3.89
13239	13240	2	ORF1ab	3.79
24911	24912	2	S	3.65

Table S3: The 10 most diverse consecutive regions in the genome in the public data, ranked by their average entropy.

start	end	length	gene	log average entropy
508	523	16	ORF1ab	5.46
29187	29188	2	N	5.05
26775	26800	26	M	5.02
23550	23583	34	S	4.62
4529	4536	8	ORF1ab	3.56
24013	24028	16	S	3.51
28253	28254	2	ORF8	3.37
4542	4556	15	ORF1ab	3.25
13454	13494	41	ORF1ab	3.03
28881	28883	3	N	2.98

Table S4: The 10 most diverse consecutive regions in the genome in the Swiss data, ranked by their average entropy.

sample ID	log entropy	mutated positions (%)	sample ID	log entropy	mutated positions (%)
SRR11577862	7.08	17.41	SRR11578392	6.82	18.40
SRR11578151	7.03	29.11	SRR11578356	6.76	19.23
SRR11577865	6.97	19.92	SRR11578031	6.67	11.54
SRR11578029	6.90	14.60	SRR11577858	6.61	17.94
SRR11578028	6.88	20.05	SRR11578137	6.54	19.26

Table S5: The 10 samples with the highest measured diversity, ranked by their entropy, and the fraction of their positions affected by mutations.
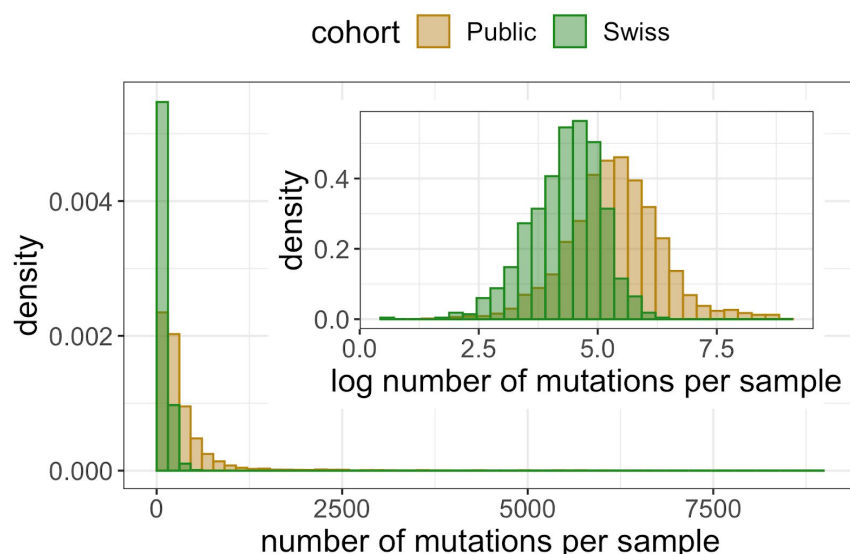

Figure S3: The distributions of the number of mutated positions per sample. Inset: the distribution under a logarithmic transform.
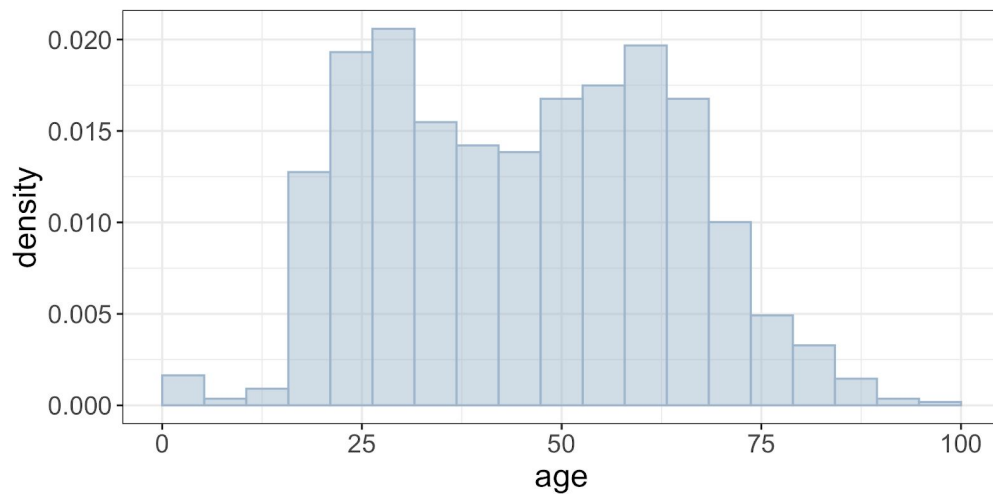
Figure S4: The distributions of ages of the 1043 public samples in the regression modelling.
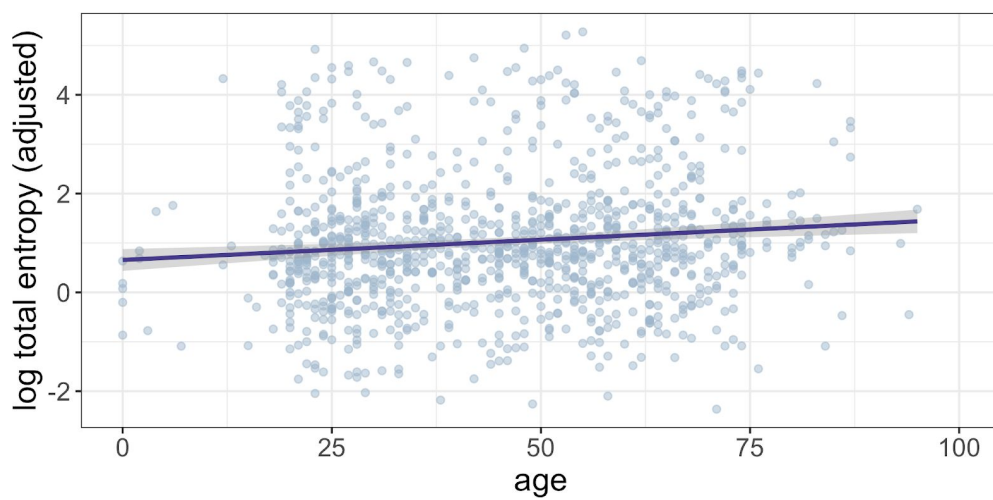

Figure S5: The dependence of log total entropy on age for the 1043 public samples, after adjustment for sex and sequencing coverage covariates in the regression modelling.
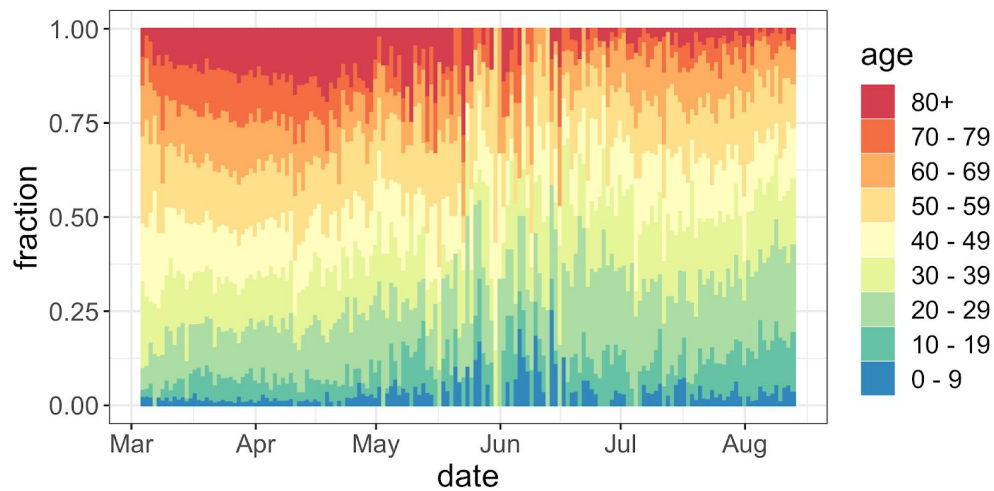
Figure S6: The distribution of ages of positive Covid-19 tests in Switzerland as a whole, covering the period where our cohort of deeply sequenced samples was collected.
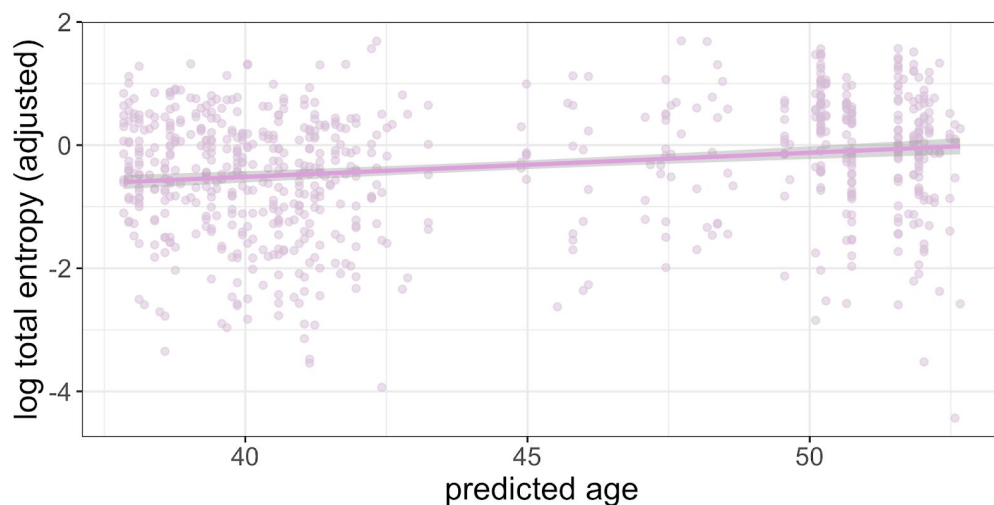

Figure S7: The dependence of log total entropy on predicted age for the 749 Swiss samples, after adjustment for sequencing coverage covariates in the regression modelling. The predicted age is constructed from a linear model of age on date built with the data displayed in Figure S6.