

Supplementary Material:

Ultra-deep sequencing reveals intra-host diversity and selective constraint in SARS-CoV-2 samples from Chile

Carol Moraga^{1,2}, Francisco Kirhman^{1,2}, Barbara Bernal^{1,2}, Sofia Poblete^{1,2}, Pastor Jullian^{1,2}, Adrian Gonzalez⁴, Mauricio Latorre^{3,5}, Alex Di Genova^{1,2,3†}

¹ Computational biology Lab, Instituto de Ciencias de la Ingeniería, Universidad de O'Higgins, Rancagua, Chile

² Centro-UOH de BioIngeniería (CUBI), Universidad de O'Higgins, Rancagua, Chile.

³ Centro de Modelamiento Matemático (CMM) UMI-CNRS 2807, Universidad de Chile, Santiago, Chile.

⁴ Tecnología, Ciencia, Laboratorio, TCL. Santiago, Chile.

⁵ Laboratorio de Bioingeniería, Universidad de O'Higgins, Rancagua, Chile.

[†] Corresponding author.

Supplementary Table 1: Lineage-specific mutation support for candidate SARS-CoV-2 co-infections. Candidate co-infections were identified using a lineage-specific mutation-matching method. For each candidate, we report the sample identifier, major lineage, number of supported major-lineage-defining sites, fraction of retained major-lineage sites supported, candidate minor lineage, number of minor-lineage-specific supported sites, median allele frequency of the minor-lineage-specific mutations, confidence tier, and the individual mutations supporting the minor lineage.

sample	major lineage	major supported sites	major frac supported sites	minor lineage	minor supported sites	minor median af	confidence tier	minor supported mutations
32	C.37	13	0.93	AY.1	5	0.37	low	22813:G>T 24410:G>A 25469:C>T 27638:T>C 28461:A>G
83	C.37	12	0.86	AY.1	5	0.36	low	22813:G>T 24410:G>A 25469:C>T 27638:T>C 28461:A>G
36	B.1.1.7	17	0.81	AY.1	5	0.36	low	22813:G>T 24410:G>A 25469:C>T 27638:T>C 28461:A>G
90	C.37	12	0.86	AY.1	5	0.36	low	22813:G>T 24410:G>A 25469:C>T 27638:T>C 28461:A>G
96	B.1.1.529	20	0.69	AY.1	5	0.29	low	24410:G>A 25469:C>T 27638:T>C 28461:A>G 29402:G>T
43	P.1	9	0.64	AY.1	5	0.23	low	22813:G>T 22995:C>A 24410:G>A 25469:C>T 28461:A>G
45	P.1	9	0.64	B.1.1.529	5	0.17	low	22992:G>A 22995:C>A 23013:A>C 23948:G>T 28311:C>T
30	P.1	8	0.57	B.1.1.529	5	0.14	low	22992:G>A 22995:C>A 23013:A>C 23948:G>T 28311:C>T
40	P.1	8	0.57	AY.1	7	0.24	medium	22813:G>T 22995:C>A 24410:G>A 25469:C>T 27638:T>C 28461:A>G 29402:G>T
59	B.1.1.529	18	0.62	C.37	7	0.13	medium	10097:G>A 21786:G>T 21789:C>T 23031:T>C 24138:C>A 7424:T>G 9867:T>C
44	P.1	9	0.64	AY.1	6	0.31	medium	22813:G>T 22995:C>A 24410:G>A 25469:C>T 27638:T>C 28461:A>G
25	P.1	8	0.57	AY.1	6	0.26	medium	22813:G>T 22995:C>A 25469:C>T 27638:T>C 28461:A>G 29402:G>T
75	C.37	12	0.86	B.1.1.529	6	0.14	medium	10449:C>A 22992:G>A 23013:A>C 23948:G>T 26709:G>A 27807:C>T

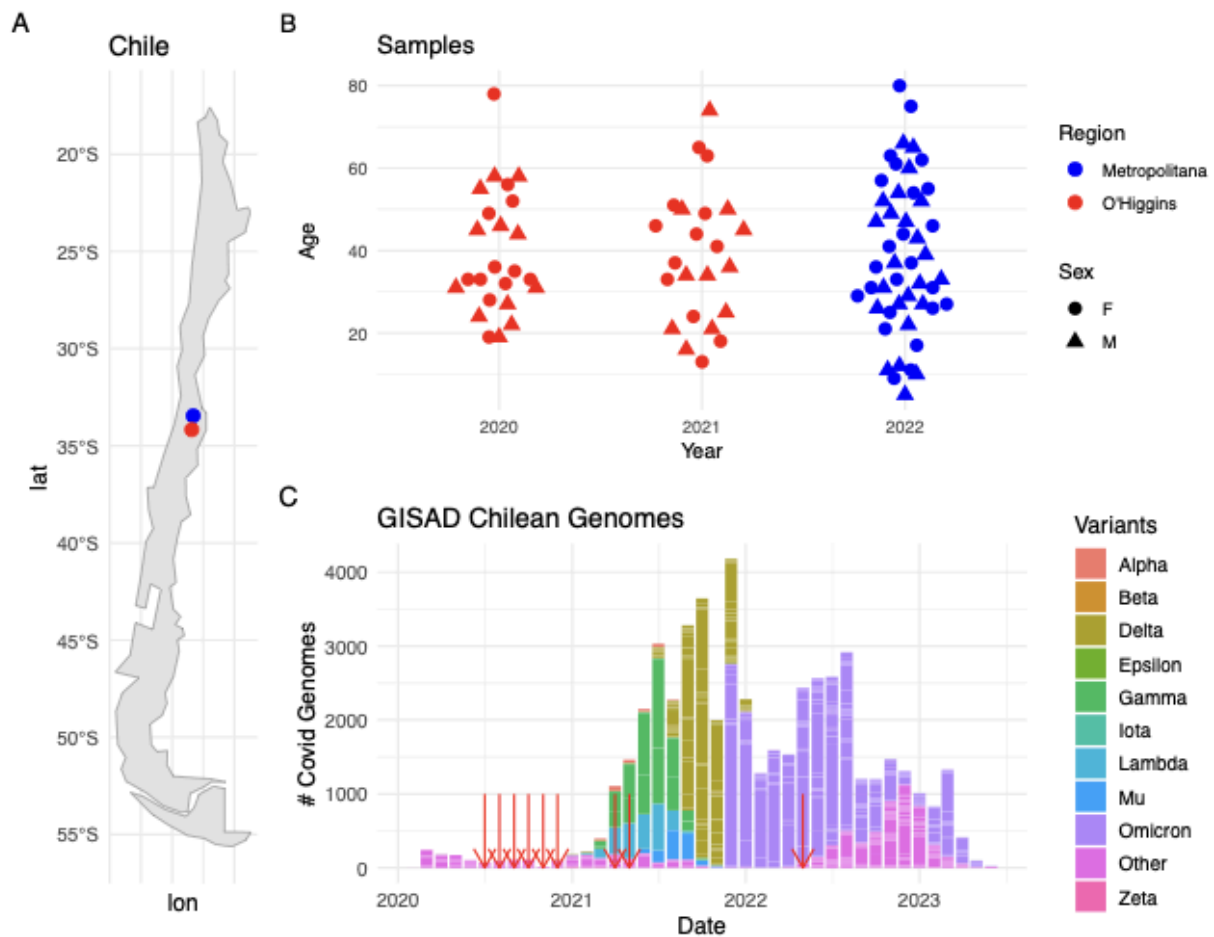


Figure S1: Overview of COVID-19 Sample Sequencing. A) Geographic Origin of Samples: The samples were collected from two regions in Chile: O'Higgins and the Metropolitan capital, Santiago. B) Demographic and Temporal Distribution of samples. C) Integration with GISAD Data: The red arrows mark the collection dates of our sequenced samples against the backdrop of Chilean genomes submitted to GISAID from 2020 to 2023 (45466 genomes at 22-09-23). This juxtaposition highlights the contribution of our data to Chilean surveillance efforts.

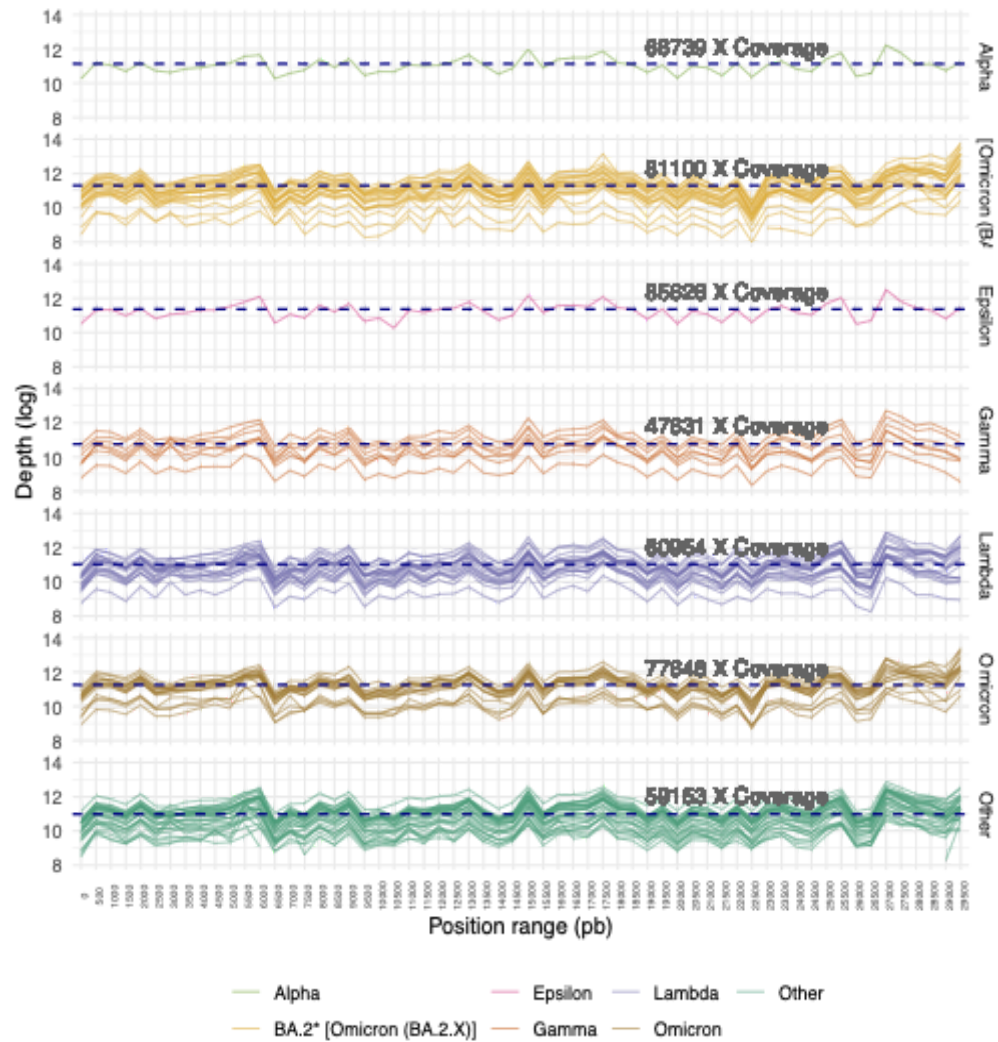


Figure S2. Sequencing depth of samples by the position of the genome. Samples are separated by variant (colors) and patient (multiple curves). Depth value is represented as media in windows of 0.5K pb.

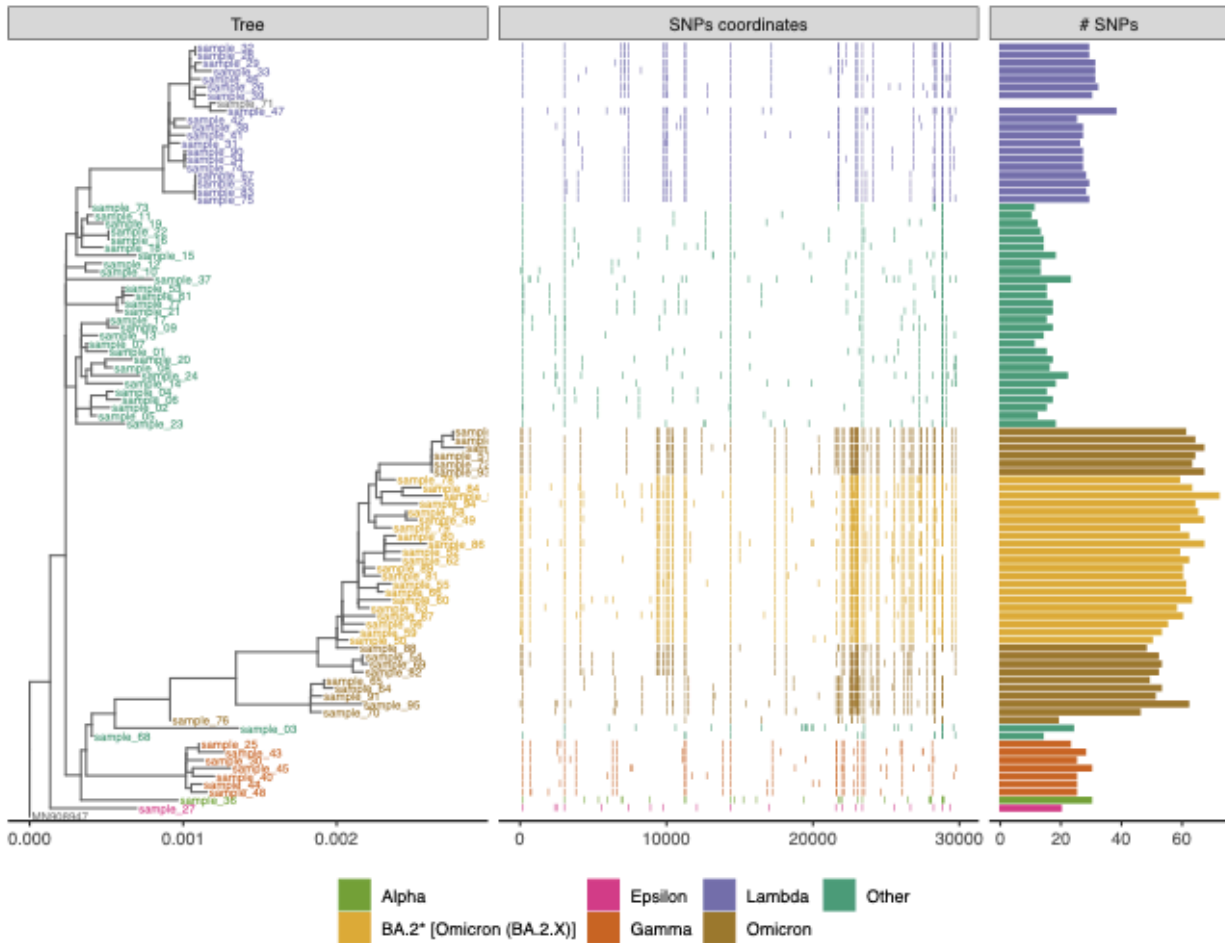


Figure S3. Phylogeny, genome-wide SNP distribution, and mutational burden of Chilean SARS-CoV-2 genomes. The maximum-likelihood tree (left) separates Alpha, Gamma, Lambda, Omicron (BA.2*), Epsilon, and Other lineages. The central panel shows SNP positions along the SARS-CoV-2 genome for each sample, ordered by the phylogeny. The right panel shows total SNP counts, with Omicron exhibiting the greatest divergence from earlier variants.

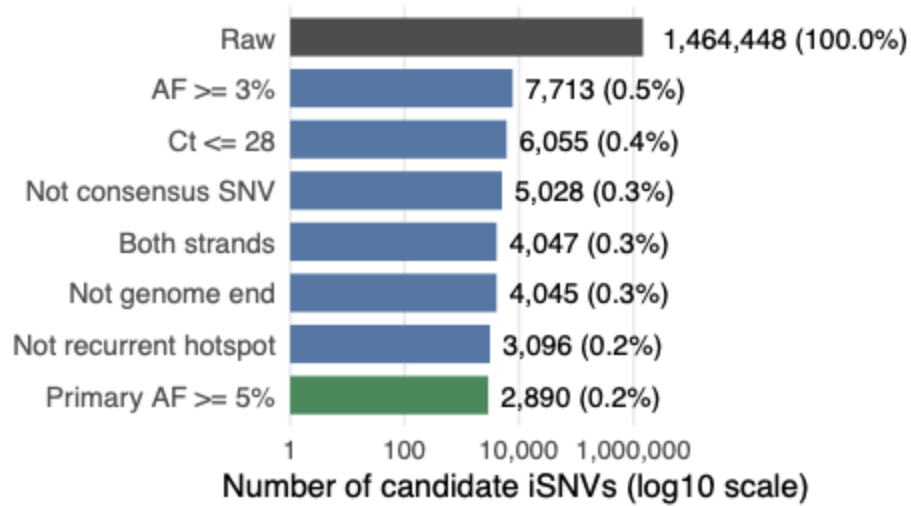


Figure S4. Filtering steps for iSNVs. Starting from 1,464,448 raw candidate iSNV variants, calls were sequentially filtered by allele frequency, sample Ct value, removal of consensus-level SNVs, strand support, genome-end masking, and exclusion of recurrent low-frequency hotspots. The final high-confidence primary iSNV set included 2,890 variants with allele frequency $\geq 5\%$, corresponding to 0.2% of the raw candidate set. Finally, the sensitive iSNV set includes 3,096 variants with $AF \geq 3\%$, corresponding to 0.21% of the raw set.

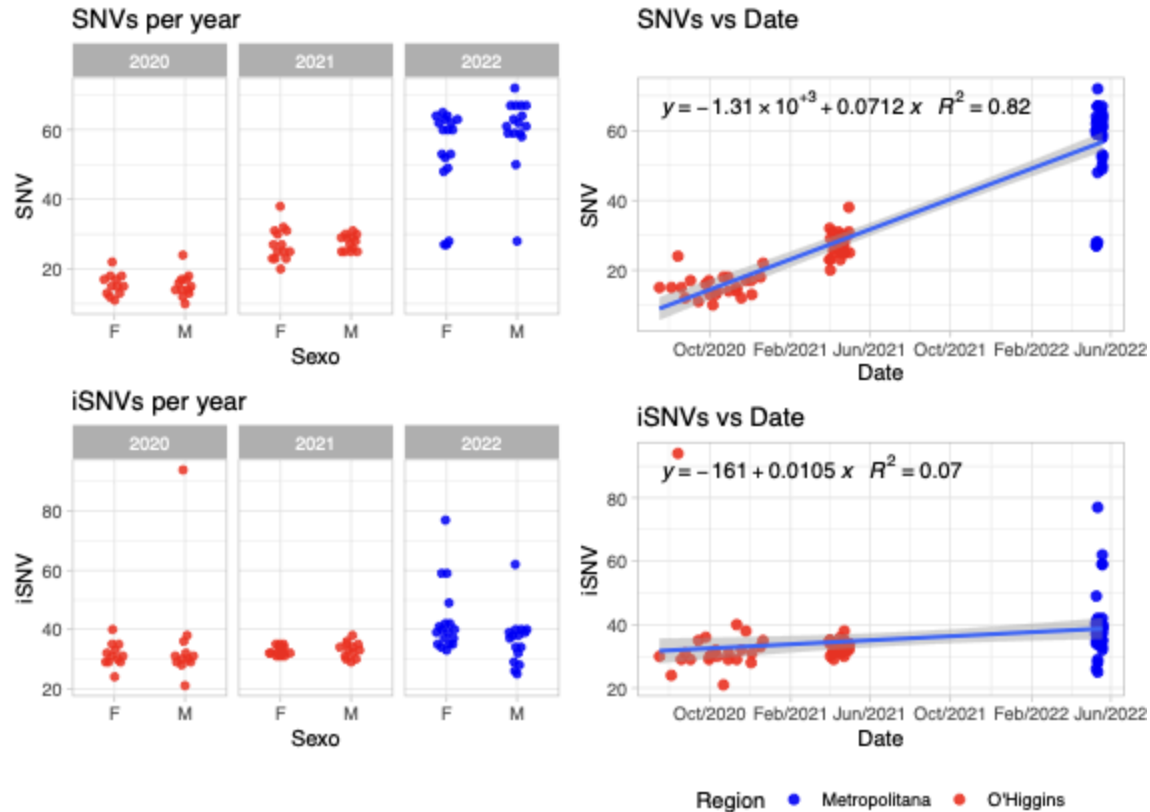


Figure S5. Single Nucleotide Variants (SNVs) and intra-Host Single Nucleotide Variants (iSNVs) comparison by Year, Sex, and Region of Collection. The top panels show SNVs per year (2020–2022) categorized by sex (F: Female, M: Male) with data from two regions: Metropolitana (blue) and O'Higgins (red), alongside a regression analysis revealing a strong positive correlation between SNVs and collection date ($R^2 = 0.82$). The bottom panels present iSNVs per year under the same categories, with a weaker correlation between iSNVs and collection date ($R^2 = 0.07$).