Appendix

Title: Genomic epidemiology of SARS-CoV-2 in Colombia

This PDF file includes:

Supplementary Text

Figs. S1 to S2

Other Supplementary Materials for this manuscript include the following:

Table S1. GISAID's nCoV-19 Acknowledgements.

Table S2. Nucleotide substitution patterns of the different lineages of SARS-CoV-2 circulating in Colombia.

Table S3. Amino acid change patterns of the different lineages of SARS-CoV-2 circulating in Colombia.

Table S4. Estimates of evolutionary divergence of SARS-CoV-2 over sequence pairs within and between lineages/sublineages from Colombia.Figs. S1 to S2.

Supplementary Text

Potential routes of COVID-19 importation in Colombia

For each air travel route (o, d) where o is a country of origin and d a destination, we calculate the proportion $E_{o,d}$ of expected importations along route (o, d) using the incidence i_O for the country o and the total number $p_{o,d}$ of passengers on route (o, d):

$$E_{o,\,d} = 100\,\frac{i_o\, p_{o,\,d}}{\sum\limits_{(u,t)} i_u\, p_{u,\,t}}$$

where the sums are taken over all possible routes. Therefore, the proportion of expect E_d importations arriving to is

$$E_d = 100\,\frac{\sum\limits_{u} i_u\, p_{u,\,d}}{\sum\limits_{(u,t)} i_u\, p_{u,\,t}}$$

and the proportion of expected importations departing from is

$$E_o = 100\,\frac{\sum\limits_{t} i_o\, p_{o,\,t}}{\sum\limits_{(u,t)} i_u\, p_{u,\,t}}$$

Epidemiological investigation per lineage

Lineage A.1.2 was detected in the Manizales city from an imported case arriving on March 3 from the United States. Lineage A.2 was identified in the Anserma municipality from a

sample collected on March 26 according to the previously published data available at GISAID. Associated traveling and contact-tracing history was not available. Lineage A.5 was identified in a transmission chain of three individuals in the Medellin city without travel history, the first one presenting symptoms on March 9. These sequences shared a distinctive substitution pattern at the amino acid (G238C at the Nucleocapsid) and nucleotide (C17470T at ORF1ab (Helicase)) levels (Tables S2 and S3). A fourth unrelated case from this lineage was identified in the Ibague city from an imported case arriving on March 17 from an unknown country. Two independent introductions of the SARS-CoV-2 lineage A.5 could explain their current epidemiology in Colombia.

The lineage B was assigned to some SARS-CoV-2 sequences whose genetic variability did not allow the assignment to a specific sublineage. A first patient (Colombia/Cali/79449) arrived in the Cali city from Spain on March 7. The viral sequence displayed a very similar pattern to lineage B.2, with two amino acid changes, L3606F and G251V at ORF1ab (Nsp6) and ORF3a, respectively (Table S3). The other two patients were identified in the same urban area (Medellin and Sabaneta municipalities belong to the Aburra Valley), with the first one presenting symptoms on March 13. These results suggest two independent introduction events to explain the current presence and distribution of SARS-CoV-2 basal lineage B in Colombia.

Lineage B.1 represented the 48 % of the analyzed SARS-CoV-2 sequences from Colombia. The genomic analysis of the genetic variability at the amino acid level revealed 4 substitution patterns (Table S3), which increased to 5 when analyzed at the nucleotide level (Table S2). The first pattern was identified in two sequences from Neiva city belonging to the

same transmission chain without travel history. The second one was identified mainly in 8 municipalities from the Valle del Cauca department, with most sequences from the capital city (Cali). These sequences displayed a distinctive amino acid change (T265I) at ORF1ab (Nsp2) (Table S3). Eight of the patients referred to international travel history from Spain and the USA, the earliest arriving on March 6, 2020. The third one was mainly distributed in the Caribbean region (Cienaga, Monteria, Santa Marta, and Valledupar) and municipalities from the center of the country near the Capital, the first case belonging to this pattern corresponded to an imported case entering the country on March 1, from Spain, early during the pandemics in Colombia and previous to the domestic and international flights restriction. The forth one was identified in several municipalities without geographic proximity, such as Medellin, Cartagena, Ibague, Leticia and Togui (this case referred a recent travel to a national tourist region), from confirmed cases without international travel history from the available data, the first case being identified in Cartagena city on March 10, 2020. Finally, the fifth pattern was identified in three sequences from samples collected in Neiva and Envigado cities, corresponding to different regions of the country, with the first patient presenting symptoms on March 18 from, two days after entering the country from Panama. These sequences displayed a distinctive amino acid change (I71V) at ORF1ab (Leader). Patterns 2 to 5 shared the amino acid change Q57H at ORF3a, while patterns 3 to 5 shared the synonymous substitution C18877T (3'-to-5' Exonuclease) and patterns 4 and 5 shared the synonymous substitution C10509T (Nsp5A) (Tables S2 and S3).

Lineage B.1.1 was not directly assigned by PANGOLIN, however it was manually defined for those sequences classified as B.1 that possessed the substitutions G28881A,G28882A,G28883C at the nucleocapsid gene (leading to the amino acid changes

R203K and G204R (Table S3) (https://github.com/hCoV-2019/lineages) and corroborated through GISAID assignment. This lineage was represented by 16 sequences, including the first confirmed case of COVID-19 entering the country on February 26 from Italy and displays five nucleotide substitutions common to pairs of sequences, four of which also displayed geographic correspondence, and therefore representing previously identified or unidentified local transmission chains. Two sequences obtained from patients from Villavicencio city without travelling history belonged to a previously characterized transmission chain and displayed two distinctive amino acid changes, S2488F and T14A at Orf1ab (Nsp3) and ORF7a, respectively (Table S3).

Lineage B.1.3 was identified in the Pereira city and Dosquebradas municipality (belonging to the same metropolitan area) from patients without traveling history, the first one presenting symptoms on April 2. Despite there is not available information about their epidemiologic relationship, the fact they shared the same nucleotide substitution pattern, geographic distribution and temporality suggest that they belong to the same transmission cluster.

Lineage B.1.5 was the second in frequency, accounting for the 26% of the analyzed sequences from Colombia. While any amino acid change is distinctive of this lineage, all the sequences shared the exclusive synonymous substitution A20268G at ORF1ab (EndoRNAse). The analysis of the genetic variability at the amino acid level revealed 3 substitution patterns (Table S3), which increased to 5 when analyzed at the nucleotide level (Table S2). The first substitution pattern was represented by sequences from Barranquilla, Bello, Bogota, Cali,

Cucuta, Medellin, Neiva, Pacho, Palmira and Tierralta municipalities, located in seven different departments. Despite the lack of geographic clustering, 12 of the 18 patients belonging to this group referred to have entered the country from Spain, France or Italy, and more interestingly, 5 patients entered on March 9, 2020 and 3 entered on March 12, 2020. The second pattern was defined by the presence of the synonymous substitution C23443T at Spike gene. According to the available information from two of the patients, no travel history was referred. Pattern 3 sequences shared the G29734C substitution at the 3'UTR and was distributed in Pereira, Cali and Yumbo, belonging to the interconnected departments Valle del Cauca and Risaralda. The first patient referred to travel history entering the country from Spain also on March 9, 2020 and referring symptoms on the same day. The fourth pattern was identified in Cali city from two patients belonging to the same transmission chain who shared the amino acid change R191C at the Nucleocapsid protein and the synonymous substitution C1327T at Orf1ab (Nsp2). The fifth substitution pattern was defined by two distinctive amino acid changes at Orf1ab, A3610V and G5063S at the Nsp6 and RdRp genes, respectively. This pattern was identified in Cali city and Buga municipality, both in the department of Valle del Cauca, the first patient referring travel history from Spain on March 13, 2020.

Lineage B.1.8 was identified in the Pereira city from a single sequence available at GISAID obtained from a patient arriving from Spain on March 15. Lineage B.1.11 was identified in Itagui (Aburra Valley) from a patient presenting symptoms on March 23. Lineage B.2 was identified in the Cali city from a single sequence obtained from a patient presenting symptoms on March 29,

without travel history. Finally, lineage B.2.5 was represented by a previously referred transmission chain with the first patient arriving in Armenia on March 10, from Italy.

Evolutionary divergence within and between lineages

The estimate of average evolutionary divergence between sequence pairs of lineages A and B was 0,000401 base substitutions per site, while those estimates for within each lineage sequence pairs were 0,000169 and 0,000222 for lineages A and B, respectively (Table S4). For a better resolution in the estimation of evolutionary divergence within and between groups, sequences were grouped according to different sublineage levels. The analysis at sublineages level 1 allowed the sequences to be grouped as belonging to A.x and B.x, while at sublineage level 2 as belonging to A.x.x and B.x.x. These comparisons at sublineage level 1 allowed to identify a higher estimated evolutionary divergence within sublineages B.1 and B.2. The estimates of genetic distance between sublineages was higher for comparisons of A.2 with the other sublineages. Interestingly, the average divergence between sublineages assigned to the same lineage (A.1 vs. A.2, A.2 vs. A.5) was higher than other comparisons between sublineages assigned to different lineages (e.g. A.1 vs. B.1, A.1 vs. B Basal, A.2 vs. B Basal). The higher resolution comparisons at sublineages level 2 allowed to identify the higher within sublineage divergence for B.1.1 and B Basal. The estimates of evolutionary divergence between sublineages showed higher values when A.2 Basal was compared to any sublineage from the B lineage and also when B.2.5 was compared to any other sublineage. The average divergence between sublineages assigned to the same lineage (A.1.2 vs. A.2 Basal, A.2 Basal vs. A.5 Basal, B.1.1 vs. B.2.5, B.1.3 vs. B.2.5) was higher than other comparisons between sublineages assigned to

different lineages (e.g. A.1.2 vs. B.1.5, B.1.8, B.1.11, B.1 Basal, B.2 Basal, B Basal) (Table S4). While these results can be affected by the low sequence number for some lineages, the estimates of the evolutionary distance between groups are in agreement with the lineage assignment and with the substitution pattern observed to be unique to every lineage (Tables S3 and S4).

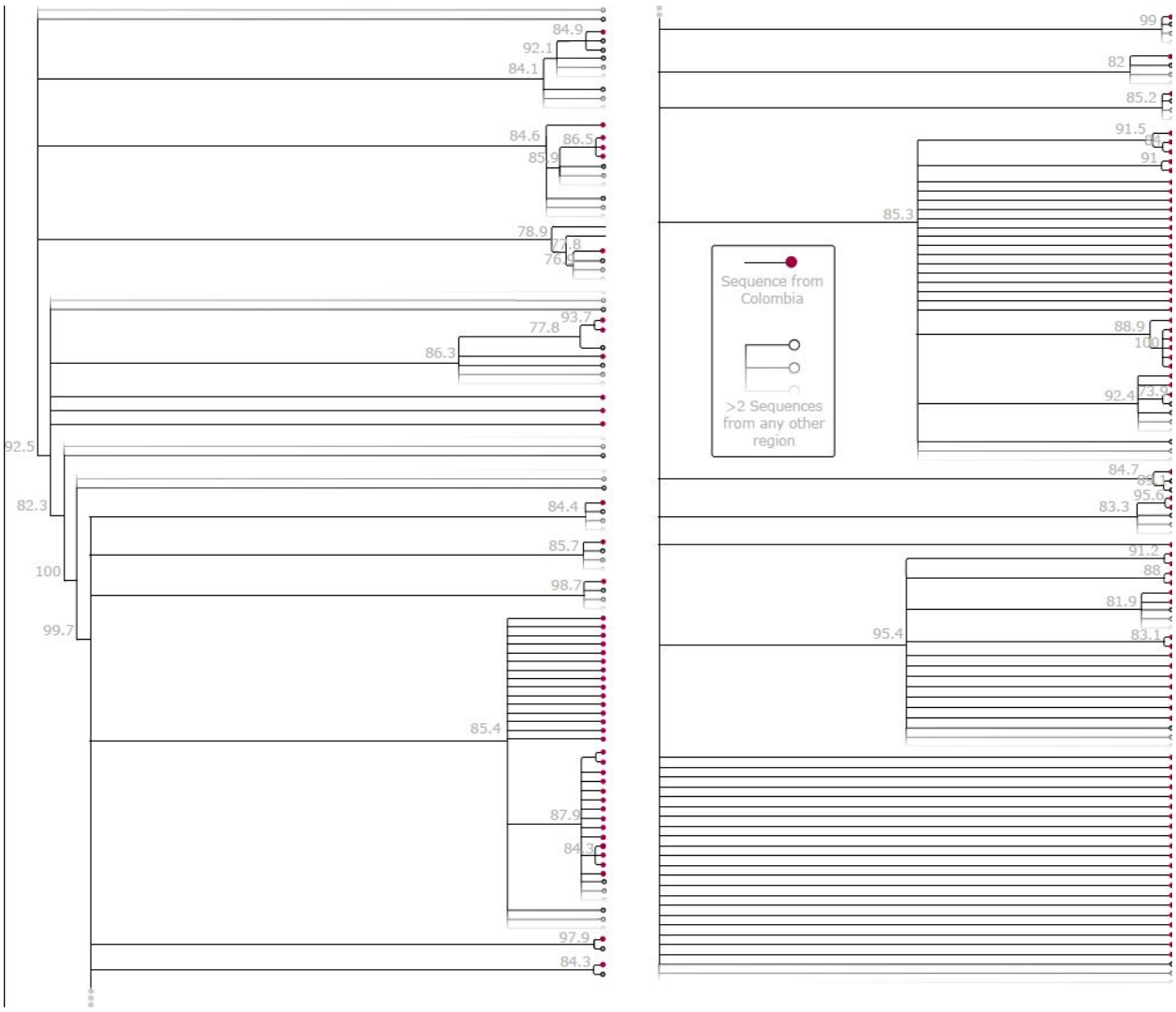

Figure S1. Schematic of the maximum likelihood tree of 1578 SARS-CoV-2 sequences. Branches with branch supports (approximate likelihood ratio test [aLRT]) below 0.7 were collapsed. Sequences from Colombia are indicated in red. Sequences from other regions in polytomies were reduced to a maximum of three sequences and represented as fading taxa.

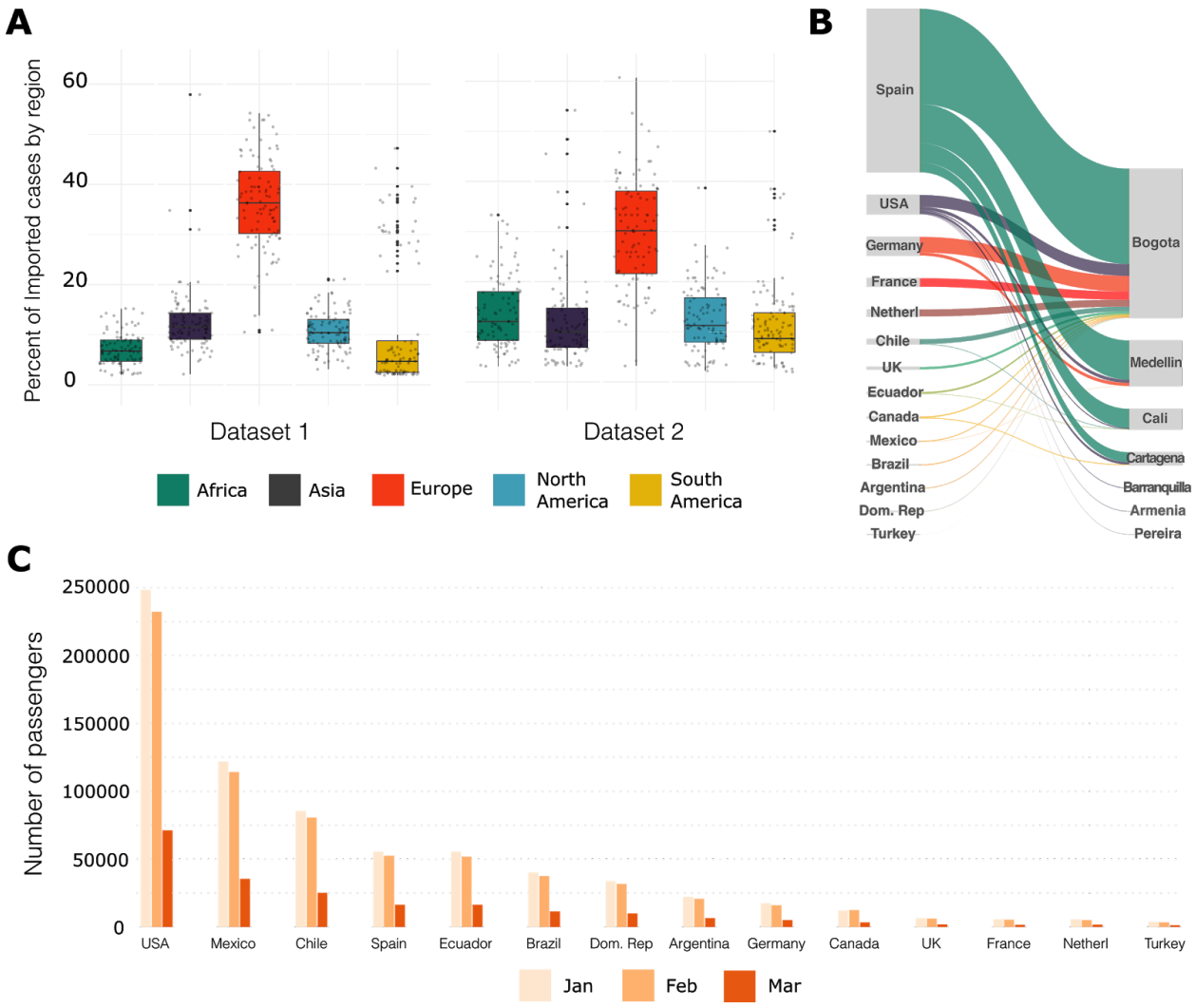

Figure S2. Potential countries of introductions of SARS-CoV-2

A) Geographical source attribution for every transition into Colombia derived from the migration inference retaining a number of sequences per region (whenever possible) equal to the number of sequences available for Colombia and retaining 50 sequences per region and all sequences from Colombia. B) Number of passengers arriving in Colombia from different countries. C) Geographical contribution inferred using Air travel data per country.