Genomic Epidemiology of Severe Acute Respiratory Syndrome Coronavirus 2, Colombia

Katherine Laiton-Donato, Christian Julián Villabona-Arenas, José A. Usme-Ciro, Carlos Franco-Muñoz, Diego A. Álvarez-Díaz, Liz Stephany Villabona-Arenas, Susy Echeverría-Londoño, Zulma M. Cucunubá, Nicolás D. Franco-Sierra, Astrid C. Flórez, Carolina Ferro, Nadim J. Ajami, Diana Marcela Walteros, Franklin Prieto, Carlos Andrés Durán, Martha Lucia Ospina-Martínez, Marcela Mercado-Reyes

Coronavirus disease (COVID-19) in Colombia was first diagnosed in a traveler arriving from Italy on February 26, 2020. However, limited data are available on the origins and number of introductions of COVID-19 into the country. We sequenced the causative agent of COVID-19, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), from 43 clinical samples we collected, along with another 79 genome sequences available from Colombia. We investigated the emergence and importation routes for SARS-CoV-2 into Colombia by using epidemiologic, historical air travel, and phylogenetic observations. Our study provides evidence of multiple introductions, mostly from Europe, and documents ≥12 lineages. Phylogenetic findings validate the lineage diversity, support multiple importation events, and demonstrate the evolutionary relationship of epidemiologically linked transmission chains. Our results reconstruct the early evolutionary history of SARS-CoV-2 in Colombia and highlight the advantages of genome sequencing to complement COVID-19 outbreak investigations.

Coronavirus disease (COVID-19) is a life-threatening respiratory illness caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), an emerging zoonotic virus first identified in Wuhan, China (*1*). The first confirmed cases of COVID-19 were reported on January 12, 2020, from patients who had respiratory symptoms during December 8, 2019–January 2, 2020 (*2*). Despite early containment and mitigation measures (*3*), the high infectiousness, presymptomatic transmission, and prolonged transmissibility of SARS-CoV-2 (*4*,*5*) combined with other factors, such as globalization, led to the rapid spread of COVID-19 across the world.

Rigorous contact-tracing and physical distancing measures implemented in different countries have been effective in delaying the epidemic during the contention phase (*6*–*9*). However, ensuing lockdowns and travel restrictions to minimize the burden on healthcare systems have led to a decline in wellbeing and an economic downturn and have had profound impacts in low-to-middle income countries (*10*). The contention phase in Colombia started on March 6, 2020, when the Instituto Nacional de Salud (INS; National Institute of Health) confirmed the first case of COVID-19 from a person returning to Colombia from Italy on February 26, 2020 (*11*). On March 23, a total 314 cases had been confirmed, which prompted the closure of all the country borders to contain the outbreak. On March 31, ≥10% of confirmed cases were among persons with no known exposure to a COVID-19 patient (*12*), presumably due to extensive community transmission. Colombia then implemented the mitigation phase, which included physical distancing as the main strategy to limit virus spread. By June 18, a total of 57,046 confirmed cases and 1,864 deaths had been reported in Colombia (*13*).

The unprecedented global health and societal emergency posed by the COVID-19 pandemic urged data sharing and faster-than-ever outbreak research developments that are reflected in the >37,000 complete SARS-CoV-2 genomes made available through

Author affiliations: Instituto Nacional de Salud, Bogotá, Colombia (K. Laiton-Donato, J.A. Usme-Ciro, C. Franco-Muñoz, D.A. Álvarez-Díaz, A.C. Flórez, C. Ferro, D.M. Walteros, F. Prieto, C.A. Durán, M.L. Ospina-Martínez, M. Mercado-Reyes); Centre for the Mathematical Modelling of Infectious Diseases (CMMID) and London School of Hygiene & Tropical Medicine, London, UK (C.J. Villabona-Arenas); Universidad Cooperativa de Colombia, Santa Marta, Colombia (J.A. Usme-Ciro); Universidad Industrial de Santander, Bucaramanga, Colombia (L.S. Villabona-Arenas); Imperial College-London, London, UK (S. Echeverría-Londoño, Z.M. Cucunubá); Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Colombia (N. Franco-Sierra); Baylor College of Medicine, Houston, Texas, USA (N.J. Ajami)

DOI: https://doi.org/10.3201/eid2612.202969

public databases, mainly GISAID (https://www.gisaid.org). SARS-CoV-2 is an RNA virus with an estimated substitution rate of 0.8–1.1 × 10^{-3} substitutions/site/year (S. Duchene et al., unpub data, https://www.biorxiv.org/content/10.1101/2020.05.04.077735v1; M. Worobey et al., unpub. data, https://www.biorxiv.org/content/10.1101/2020.05.21.109322v1), which means it rapidly evolves as it is transmitted. The availability of SARS-CoV-2 genomes enabled us to detect a rapidly generating variation, demonstrating that genomic epidemiology is a powerful approach for characterizing the outbreak (*14*). Genomic epidemiology relies on phylogenetic analysis and has enabled researchers across the world to detect SARS-CoV-2 emergence in humans, reveal the importation and local transmission chains not detected by travel history and traditional contact-tracing strategies, and trace the geographic spread and prevalence of strains bearing specific mutations of epidemiologic relevance (*15*–*17*; S. Dellicour et al, unpub data, https://www.biorxiv.org/content/10.1101/2020.05.05.078758v4; J.R. Fauver et al., unpub data, https://www.medrxiv.org/content/10.1101/2020.03.25.20043828v1).

Materials and Methods

Sample Collection and Preparation

Colombia is made up of 32 departments, which are groups of municipalities, and a capital district. INS received nasopharyngeal swabs samples from patients with clinical signs and symptoms of SARS-CoV-2 from departments across the country as part of the virological surveillance of COVID-19. INS performed quantitative reverse transcription PCR to diagnose suspected COVID-19 cases by using a method recommended and transferred by the Pan American Health Organization and World Health Organization (*18*). Because of scarce resources, we selected a total of 43 samples for genome sequencing that represented ≥1 of the earliest documented samples in each affected department or samples linked to transmission chains (Appendix 1 Table 1, https://wwwnc.cdc.gov/EID/article/26/12/20-2969-App1.xlsx). We performed viral RNA extraction by using the QIAamp Viral RNA Mini Kit (QIAGEN Inc., https://www.qiagen.com) or the MagNA Pure LC nucleic acid extraction system (Roche Diagnostics GmbH, https://lifescience.roche.com).

Genomic Library Preparation and Sequencing

Library preparation and sequencing were performed following the ARTIC network (https://artic.network) real-time molecular epidemiology for outbreak response protocol and by using both nanopore and next-generation sequencing technologies (*19*). We processed 10 samples by using the MinION sequencer (Oxford Nanopore Technologies, https://nanoporetech.com). We processed the remaining 33 samples by using the Nextera XT DNA library prep kit (Illumina, https://www.illumina.com) and performed sequencing by using the MiSeq Reagent Kit Version 2 and MiSeq sequencer (Illumina).

Genomic Sequence Assembly

We performed base calling on nanopore reads by using Guppy version 3.2.2 (Oxford Nanopore Technologies) and then demultiplexed and trimmed reads by using Porechop version 0.3.2_pre (*20*). We aligned processed reads against a SARS-CoV-2 reference genome (GenBank reference no. NC_045512.2) by using Burrows-Wheeler Aligner's Smith-Waterman Alignment (*21*). We performed base calling for single-nucleotide variants with a depth of ≥200× and then generated polished consensus by using Nanopolish version 0.13.2 (*22*). MiSeq reads were demultiplexed and we used fastp (*23*) to perform quality control using a Q-score threshold of 30. Processed reads were aligned against the SARS-CoV-2 reference genome, we performed bas calling for single nucleotide variants with a depth of ≥100× and generated consensus genomes by using Burrows-Wheeler Aligner's Smith-Waterman Alignment version 0.7.17 (*21*) and BBMap (*24*).

Phylogenetic Analysis of SARS-CoV-2 in Colombia

Sequence data covered the 20 affected departments and the capital district of Colombia. We collected 43 SARS-CoV-2 genome sequences from this study and 79 other sequences from Colombia deposited in GISAID. We combined the 122 sequences from Colombia with 1,461 representative genome sequences from South America–focused subsampling available from NextStrain (https://nextstrain.org) (*25*) as of May 20, 2020 (Appendix 1 Table 2) plus reference MN908947.3 from the GenBank nucleotide database (accesssion no. NC_045512). Across departments, a median of 1.5 sequences (mean 3.9; range 1–45) were available per department. We classified the full genomic dataset into lineages by using Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN) and aligned these with 10 iterative refinements by using MAFFT (*26*–*28*). We removed all alignment positions flagged as problematic for phylogenetic inference, including highly homoplasic positions and 3′ and 5′ ends (*29*). We performed maximum-likelihood phylogenetic reconstruction

on the curated alignment and a Hasegawa-Kishino-Yano plus gamma distribution 4 substitution model by using IQ-TREE (*30*,*31*). We estimated branch support by using an SH-like approximate likelihood ratio test (SH-aLRT) and considered ≥0.75 a high SH-aLRT (*32*). We removed 6 sequences from Colombia from further analysis because they had an inconsistent temporal signal in a clock analysis in TreeTime (*33*). We inferred time-scaled trees and rooted these with least-squares criteria and the evolutionary rate of $\geq 1.1 \times 10^{-3}$ substitutions/site/year estimated by S. Duchene et al. (unpub data, https://www.biorxiv.org/content/10.1101/2020.05.04.077735v1) by using TreeTime (*33*) and least-squares dating (*34*).

We considered geographic locations of sequence data, aggregated by continent except for Colombia, as discrete states, used these data for migration inference, and modeled transitions as a time reversible process by using TreeTime (*33*). We interpreted the number of state transitions into Colombia as a proxy for the minimum number of introductions.

In sensitivity analysis and to measure the effect of the SARS-CoV-2 uneven genomic representativeness across the world, we implemented 2 downsampling strategy datasets in which, based on location, the sequences were randomly resampled 100 times and the phylogenetic and migration inference was replicated. The downsampling strategies were as follows: retaining several sequences per region, when possible, equal to the number of sequences available for Colombia; or retaining 50 sequences per region and the total number of sequences from Colombia, which was the most even sampling per region for the South America–focused subsample.

Potential Routes of SARS-CoV-2 Importation into Colombia

We inferred the relative proportion of expected SARS-CoV-2 importations per country by considering COVID-19 incidence per number of international air passengers arriving in Colombia and the available flight travel. We obtained the number of international flights and number of passengers arriving during January 1–March 9, 2020 from the Special Administrative Unit of Civil Aeronautics of Colombia (Aerocivil, http://www.aerocivil.gov.co). The air travel data consists of direct flights from 14 countries to 7 main cities. We calculated COVID-19 incidence for each of the 14 countries with direct flights to Colombia by using the number of confirmed cases reported by the World Health Organization as of March 17, 2020, the date when travel restrictions started in Colombia (*35*), and the total population for each country for 2019 reported in the United Nations World Population Prospects 2019 database (*36*), as described in D.D.S. Candido, et al. (*37*) (Appendix 2, https://wwwnc.cdc.gov/EID/article/26/12/20-2969-App2.pdf).

Ethics Statement

According to the national law 9/1979, decrees 786/1990 and 2323/2006, the Instituto Nacional de Salud is the reference lab and health authority of the national network of laboratories and in cases of public health emergency or those in which scientific research for public health purposes as required, the Instituto Nacional de Salud may use the biological material for research purposes, without informed consent, which includes the anonymous disclosure of results. The information used for this study comes from secondary sources of data that were previously anonymized and do not represent a risk to the community.

Results

Epidemiologic Investigation of SARS-CoV-2 Introductions, Contact-Tracing, and Community Transmission

In Colombia, preventive isolation and monitoring for passengers arriving from China, Italy, France, and Spain started on March 10, 2020. A national health emergency was declared on March 12, and tougher measures then started to be set in place, including the closing of borders on March 17, the ban of international flights on March 20, and the ban of domestic flights on March 25. Implementations of lockdowns occurred from March 25 onward, including Resolutions 380 and 385 from the Colombian Ministry of Health and Social Protection (*38*,*39*); Decrees 412 and 457 from the Ministry of the Interior (*40*,*41*); and Decree 439 from the Ministry of Transport (*42*). Despite a massive drop in air traffic, >15,500 residents returned to Colombia through humanitarian flights during April–June (*43*). By June 1, >30,000 cases of COVID-19 had been documented in Colombia and 857 cases (2.8%) had been linked to travel abroad (Figure 1, panel A).

Most (816, 95.2%) imported cases were symptomatic. The prominent geographic sources for symptomatic cases were Spain (245 [28.6%] cases), the United States (203 [23.7%] cases), Ecuador (50 [5.8%] cases); Mexico (49 [5.7%] cases), and Brazil (41 [4.8%] cases). The other 41 imported cases were asymptomatic and were detected through contact tracing. Among asymptomatic imported cases, most (16, 39%) were imported from Spain, the United States (13, 31.7%), Brazil (3, 7.3%), and Mexico (2, 4.9%). Overall, most

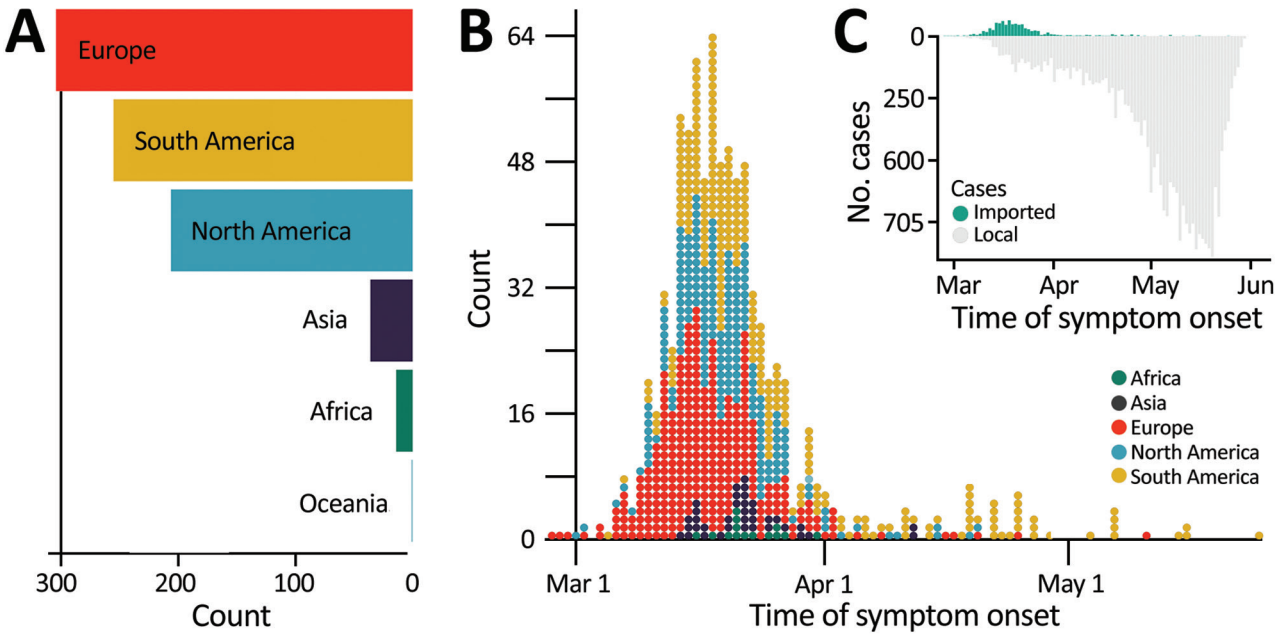


Figure 1. Proportion of imported and local cases early during the COVID-19 pandemic, Colombia. A) Region of origin for the reported imported cases. B) Distribution over time of symptomatic imported and local cases, by region of origin. C) Number of local and imported COVID-19 cases over time. COVID-19, coronavirus disease.

imported cases were from Spain (30.5%), the United States (25.2%), Mexico (6%), Ecuador (5.8%), and Brazil (5.1%). Most symptomatic imported cases were traced back to countries in Europe and the Americas.

The number of symptomatic imported cases steadily increased and peaked on March 14, when local cases were on the rise, but before border closures and the international air travel ban. Our estimate is based on the average incubation time of COVID-19 (*44*) and symptom onset but is 4.8 days earlier than the actual peak on March 18 (Figure 1, panel B). Initial introductions were predominantly linked to Europe; however, both Europe and the Americas were prominent geographic sources of infections during the onset of the epidemic. The introductions after the peak mainly occurred from countries in South America.

SARS-CoV-2 Diversity

To elucidate the dynamics of SARS-CoV-2 spread into Colombia, we combined the 43 whole-genome sequences obtained in our study with sequences from Colombia deposited in GISAID, which provided a set of 122 complete genomes. Sequences from Colombia were classified into 12 sublineages: A.1.2, A.2, A.5, B, B.1, B.1.1, B.1.3, B.1.5, B.1.8, B.1.11, B.2, and B.2.5. The proportion of lineages documented in Colombia seems to reflect founder effects. For example, sublineages B.1, B.1.1, and B.1.5 were found in the early epidemiologically linked transmission chains and consistently were observed most frequently; B.1 was observed in 59 (48.4%) cases, B.1.5 in 31 (25.4%), and B.1.1 in 16 (13.1%) (Figure 2, panel A). From the South America–focused subsampling available from NextStrain, comparable findings were observed for other countries in South America (*45*,*46*), where the most frequently observed lineages were B.1 in 149 (60.8%) cases, B.1.5 in 35 (13.5%) cases, and A.5 in 14 (5.7%) cases.

On average, we identified 1 lineage per department. For instance, the number of documented lineages was highly correlated with the availability of samples (Pearson product-moment correlation coefficient [PPMCC] = 0.72; $p<0.001$) and uncorrelated with the number of local cases (PPMCC = 0.35; $p = 0.049$). We noted 5 different lineages in the departments of Valle del Cauca and Antioquia and 3 different lineages in Cundinamarca; these departments have the most populated capitals and we had more samples from them (Figure 2, panel B). We observed a moderate positive correlation between the number of lineages documented in a department and the number of imported cases (PPMCC = 0.51; $p = 0.002$).

Molecular Evolution of SARS-CoV-2 in Colombia

We identified 133 single-nucleotide variants (NVs) by using the full genome sequences from Colombia and the reference sequence (GenBank accession no. NC_045512.2). Most NVs (131; 98.5%) fell into

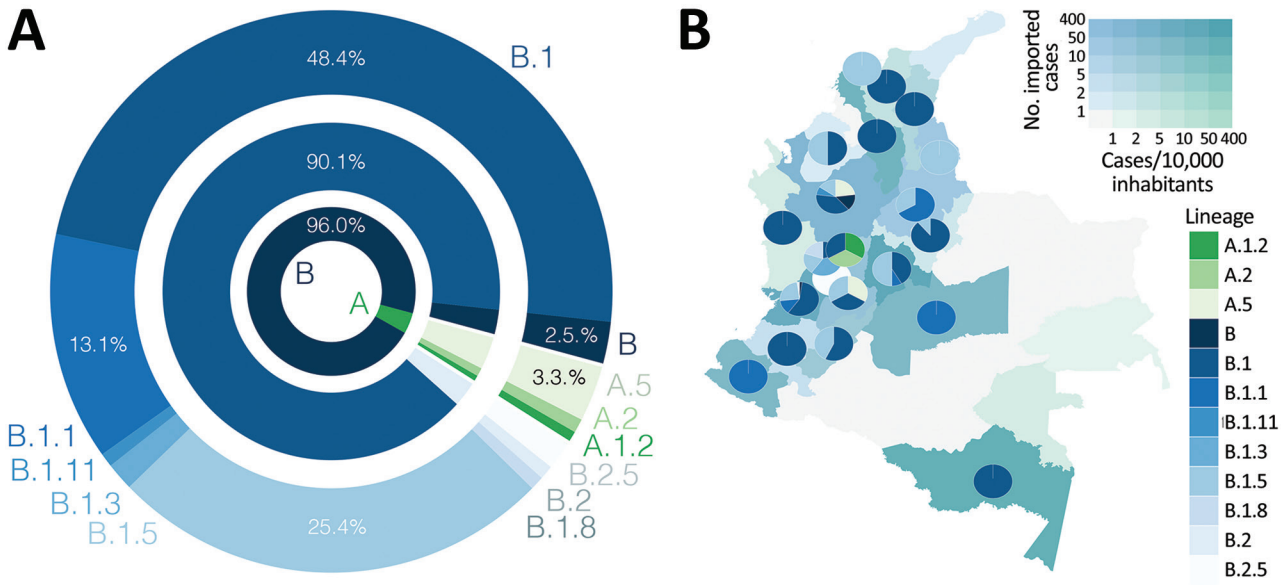


Figure 2. Frequency and distribution of SARS-CoV-2 lineages, Colombia. A) Frequency of A and B lineages and sublineages of SARS-CoV-2 identified. B) Map of distribution of lineages across the country. Departments are colored by the number of imported cases/10,000 inhabitants (inset) and the number of reported introductions. SARS-Co-V-2, severe acute respiratory syndrome coronavirus 2.

the coding region, and 1 NV was identified at each noncoding end. Among NVs in coding sites, 71 (54.2%) led to nonsynonymous substitutions. Most NVs (92/133) were unique to a sequence. Among the shared NVs, 38/41 were associated with a specific lineage (Appendix 1 Tables 3, 4). These observations suggest that the substitutions are not laboratory-specific and most likely the outcome of in situ evolution, shared ancestry, or both (Appendix 2).

In our study, among sequences with complete metadata, 90% (108/120) of sequences from Colombia displayed an amino acid change in region D614G, and the remaining 10% (12 sequences) displayed a change in region D614 (Appendix 1 Table 4). G614 has been associated with higher infectivity (L. Zhang et al., unpub data, https://www.biorxiv.org/content/10.1101/2020.06.12.148726v1) and greater transmissibility with no effects on disease severity outcomes (*46*; E.M. Volz et al., unpub data, https://www.medrxiv.org/content/10.1101/2020.07.31.20166082v2). All G614 sequences also carried mutations that segregate together as described in B. Korber et al. (*47*); we identified the nucleotide substitution C241T at 5′-UTR; the synonymous substitution C3037T at open reading frame 1ab (ORF1ab), the nonstructural protein 3 encoding-gene; and a change in P4715L aa position in ORF1ab, the RNA-dependent RNA polymerase encoding gene. The presence of these and other mutations can be phenotypically and epidemiologically relevant and warrant further monitoring.

Most patients from Colombia for whom genomic sequences were available were symptomatic (n = 90); 59.6% had cough and fever and the others had ≥1 symptom; 10 died, 70% of whom had underlying conditions (Appendix 1 Table 1). However, given the limited number of sequences available, we could not reliably investigate any genomic determinant of clinical outcome.

Evolutionary Relationships between Local and Global SARS-CoV-2 Isolates

The time-stamped phylogeny of 122 isolates from Colombia and 1,462 representative global SARS-CoV-2 isolates showed that the estimated time to the most recent common ancestor for the sampled sequence data is December 7, 2019 (range October 25–December 26, 2019) (Figure 3, panel A). Asia was the inferred ancestral state at the root. Both these observations are in line with the known epidemiology of the pandemic. A root-to-tip regression of genetic distance against sampling time evidenced consistent temporal signal in the sequence data (Figure 3, panel B). The isolates from Colombia were interspersed among the isolates from other countries, suggesting multiple introductions (Figure 3, panels A, C). However, considerable phylogenetic uncertainty appears along the tree and the fine-grained relationships of the isolates from Colombia could not be resolved with confidence (Appendix 2 Figure 1).

Phylogenetic uncertainty and uneven sampling made the quantification of the number of introductions

into the country challenging, let alone dating the time of the introductions. The number of state transitions into Colombia heavily relies on the number and nature of the sequences included from other locations (Figure 4, panel A). By using all sequences in the South America–focused subsampling available from NextStrain, we estimated that an average of 64 (interquartile range [IQR] 62–67) introductions into the country have occurred but this estimate gets lower as we reduce the number of samples (sensitivity analyses) from other locations, down to 22 with the most even downsampled dataset. Independent of the dataset, either the complete or the subsampled datasets, and in line with the epidemiologic information, most geographic source attributions are from Europe (Figure 4, panel B; Appendix 2 Figure 2). This observation also aligns with our estimates using travel data (Figure 4, panel C; Appendix 2 Figure 2).

During January–March 2020, a total of 7 cities in Colombia received 1,593,211 international passengers from 14 countries. Bogotá was the most concentrated city for flights, receiving around 77% of passengers; other cities included Medellín with 11%, Cartagena with 6%, and Cali with 4% of passengers. In total, 35% of international passengers started their journeys in the United States, 17% in Mexico, and 12% in Chile. However, we estimate 87% of all imported COVID-19 cases in Colombia came from Europe, 9.5% from North America, and 3.4% from South America. When stratified by country, the primary source of importation was Spain, which had 71.4% of imported cases; the United States had 8.4%, Germany had 8%, and France had 3.4% (Appendix 2 Figure 2). Our data show most (65.2%) COVID-19 cases were among travelers arriving in Bogotá; 20% were among those arriving in Medellín, and 9% among those arriving in Cali. We estimate that the Spain–Bogotá route carried 42% of the total imported cases.

Since the first COVID-19 case was identified in Colombia on February 26, 2020, contact-tracing efforts had been put in place. We obtained multiple

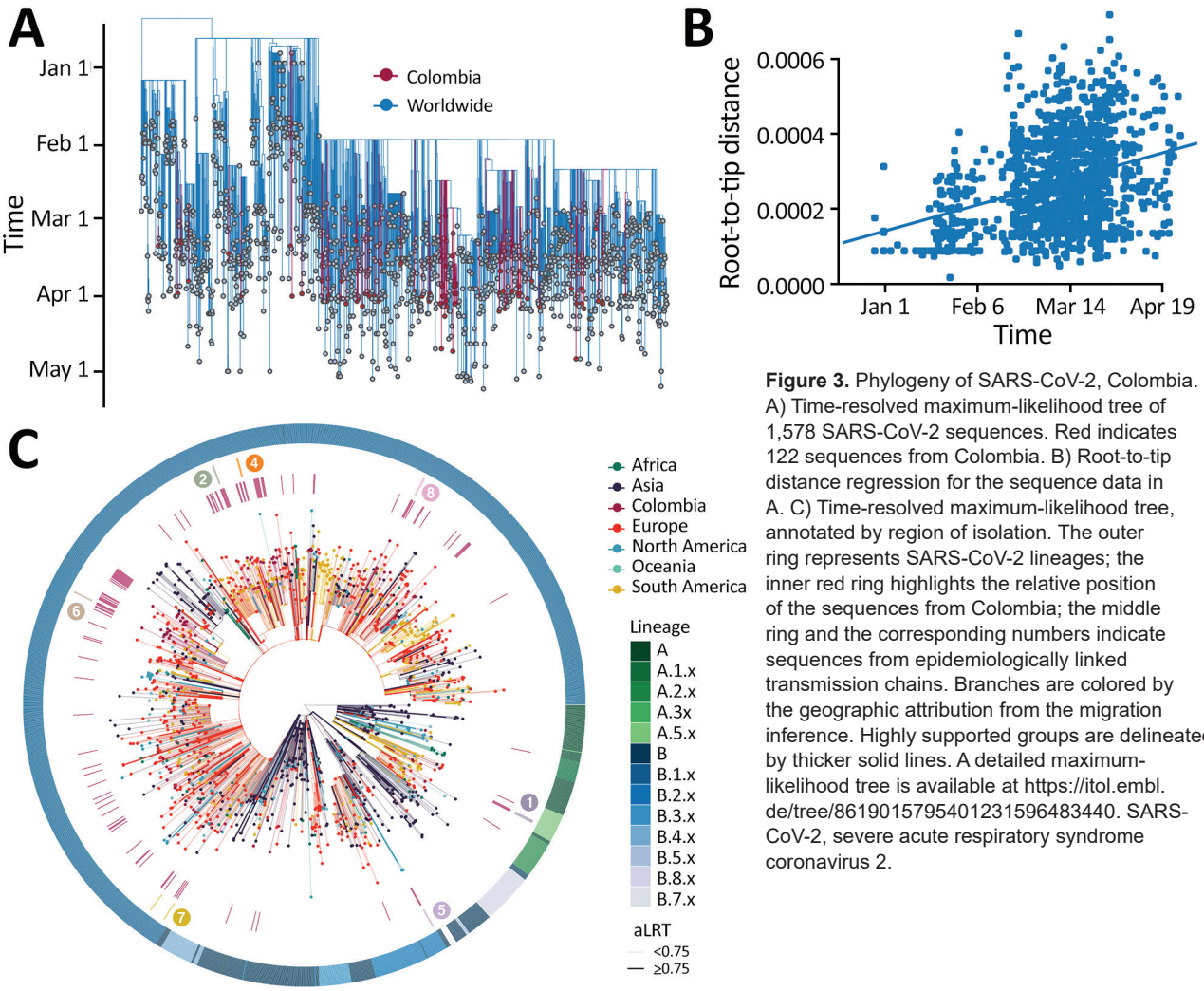


Figure 3. Phylogeny of SARS-CoV-2, Colombia. A) Time-resolved maximum-likelihood tree of 1,578 SARS-CoV-2 sequences. Red indicates 122 sequences from Colombia. B) Root-to-tip distance regression for the sequence data in A. C) Time-resolved maximum-likelihood tree, annotated by region of isolation. The outer ring represents SARS-CoV-2 lineages; the inner red ring highlights the relative position of the sequences from Colombia; the middle ring and the corresponding numbers indicate sequences from epidemiologically linked transmission chains. Branches are colored by the geographic attribution from the migration inference. Highly supported groups are delineated by thicker solid lines. A detailed maximum-likelihood tree is available at https://itol.embl.de/tree/8619015795401231596483440. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

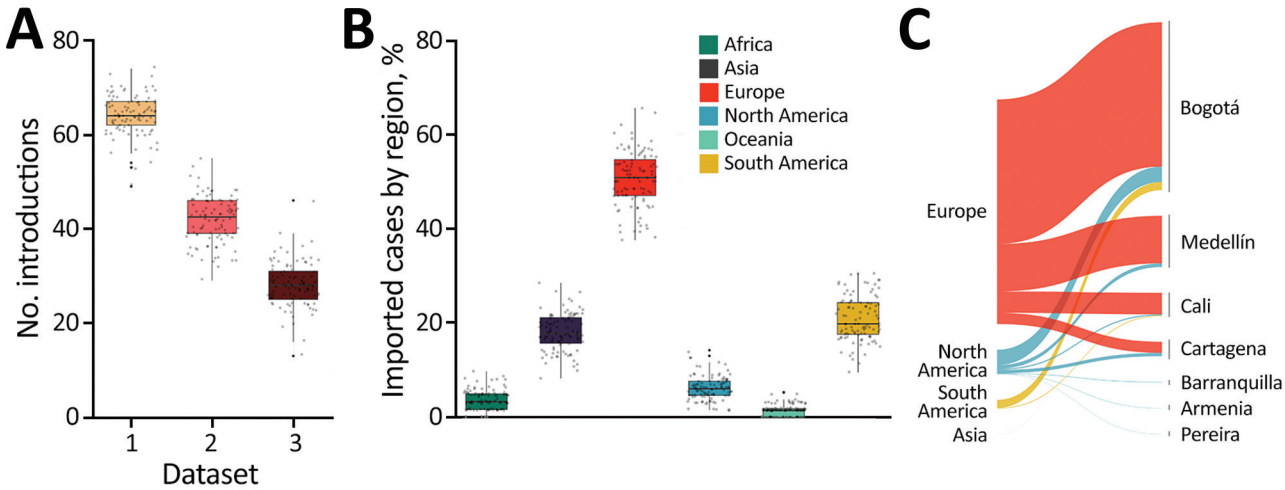


Figure 4. Potential routes of importation for SARS-CoV-2, Colombia. A) The number of transition changes into Colombia following migration inference by using all available sequences per region (dataset 1); retaining several sequences per region, when possible, equal to the number of sequences available for Colombia (dataset 2); and 50 sequences per region and all sequences from Colombia (dataset 3). Box top and bottom lines indicate 25th and 75th percentiles; horizontal lines within boxes indicate means; error bars indicate SDs. B) Geographic source attribution for every transition into Colombia derived from the migration inference using all the available sequences per region. Box top and bottom lines indicate 25th and 75th percentiles; horizontal lines within boxes indicate means; error bars indicate SDs. C) Geographic contribution inferred by using air travel data per country aggregated by region.

sequences from 7 distinct early epidemiologically linked transmission chains (Appendix 1 Table 1) and mapped these data into the phylogeny (Figure 3, panel C). All but 1 set of sequences did not group, but it appeared very close in the tree. These data underscore the potential utility of genomic epidemiology to link persons with incomplete information, such as cases that are disconnected due to intermediate asymptomatic carriers, and complement outbreak transmission investigations.

Our study has some limitations. First, the geographic sources of infection relied on persons self-reporting symptom onset and travel histories, which are subject to inaccuracies. Second, we used air travel data from likely destinations in Colombia, but other locations also might have fueled COVID-19 emergence and dissemination in the country; flight travel data was not available for dates after March 9, 2020. Third, the number of sequences sampled represented a tiny fraction of the documented number of imported cases into Colombia. The sample was selected as a countrywide representation, given limited resources for genome sequencing; thus, the introduced viral diversity also might have been underestimated. Another limitation is the inherent uncertainty stemming from global unsystematic sampling. Therefore, the inferences about the number of introductions and the corresponding geographic sources should be interpreted with caution. We attempted to overcome this by undertaking sensitivity analyses and contrasting the results with the available epidemiologic data and our estimates from travel data. However, more sequence data from Colombia and undersampled countries, together with information of sampling representativeness per country, are needed to account for sampling uncertainty in a more statistically rigorous manner.

Discussion

We describe the complete genome sequences of SARS-CoV-2 from 43 clinical samples, results of an epidemiologic investigation of imported cases, and the phylogenetic findings of 122 genome sequences from Colombia that characterize the epidemic onset of COVID-19 in the country. Our study provides evidence that several independent COVID-19 introductions occurred in Colombia and documents ≥12 SARS-CoV-2 lineages. Most of the notified introductions to countries in Latin America occurred from Europe, an observation that was supported by phylogenetic and air travel data (*48*; C. Salazar et al., unpub data, https://www.biorxiv.org/content/10.1101/2020.05.09.086223v1). Although the sequence data do not represent the actual number of epidemiologically linked transmission chains, our phylogenetic findings validated the linkage for epidemiologically linked transmission chains with available sequence data. Our results further underscore the advantages of genome sequencing to complement COVID-19 outbreak investigations and support the need for a more comprehensive country-wide study of the epidemiology and spread of SARS-CoV-2 in Colombia.

Acknowledgments

The authors thank the National Laboratory Network and Virology Group of INS for routine virologic surveillance of SARS-CoV-2 in Colombia. We also thank all researchers who deposited genomes in GISAID's EpiCoV Database contributing to genomic diversity and phylogenetic relationship of SARS-CoV-2. Finally, we thank Maylin Gonzalez Herrera for her technical assistance.

This work was funded by the Project CEMIN-4-2020 Instituto Nacional de Salud. C.J.V.-A. is supported by an ERC European Research Council Starting Grant (award no. 757688). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

About the Author

Dr. Laiton-Donato is the head of the Sequencing and Genomics Unit, Dirección de Investigación en Salud Pública, Instituto Nacional de Salud, Colombia. Her research interests include molecular virology of emerging viruses with impact in public health.

References

1. 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;579:265–9. https://doi.org/10.1038/s41586-020-2008-3
2. World Health Organization. Novel coronavirus—China, 2020 Jan 12 [cited 2020 Jun 16]. https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china
3. Kraemer MUG, Yang C-H, Gutierrez B, Wu C-H, Klein B, Pigott DM, et al.; Open COVID-19 Data Working Group. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science. 2020;368:493–7. https://doi.org/10.1126/science.abb4218
4. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. 2020;26:672–5. https://doi.org/10.1038/s41591-020-0869-5
5. Li J, Zhang L, Liu B, Song D. Case report: viral shedding for 60 days in a woman with COVID-19. Am J Trop Med Hyg. 2020;102:1210–3. https://doi.org/10.4269/ajtmh.20-0275
6. Jarvis CI, Van Zandvoort K, Gimma A, Prem K, Klepac P, Rubin GJ, et al. CMMID COVID-19 working group. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. BMC Med. 2020;18:124. https://doi.org/10.1186/s12916-020-01597-8
7. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al.; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. Lancet Public Health. 2020;5:e261–70. https://doi.org/10.1016/S2468-2667(20)30073-6
8. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al.; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. Lancet Glob Health. 2020;8:e488–96. https://doi.org/10.1016/S2214-109X(20)30074-7
9. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368:395–400. https://doi.org/10.1126/science.aba9757
10. Ribeiro F, Leist A. Who is going to pay the price of Covid-19? Reflections about an unequal Brazil. Int J Equity Health. 2020;19:91. https://doi.org/10.1186/s12939-020-01207-2
11. Ministry of Health and Social Protection. Colombia. Minsalud confirms six new cases of coronavirus (COVID-19) in Colombia [in Spanish]. Bogotá, Colombia; Boletín de Prensa no. 057 de 2020. 2020 [cited 2020 May 24]. https://www.minsalud.gov.co/Paginas/Minsalud-confirma-seis-nuevos-casos-de-coronavirus-(COVID-19)-en-Colombia.aspx
12. National Institute of Health. Colombia. COVID-2019 in Colombia daily report 2020 March 31 [in Spanish] [cited 2020 Jun 2]. https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx
13. National Institute of Health. Colombia. COVID-2019 in Colombia daily report 2020 Jun 18 [in Spanish] [cited 2020 Jun 2]. https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx
14. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al. Tracking virus outbreaks in the twenty-first century. Nat Microbiol. 2019;4:10–9. https://doi.org/10.1038/s41564-018-0296-2
15. Lu J, du Plessis L, Liu Z, Hill V, Kang M, Lin H, et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. Cell. 2020;181:997–1003.e9. https://doi.org/10.1016/j.cell.2020.04.023
16. Eden J-S, Rockett R, Carter I, Rahman H, de Ligt J, Hadfield J, et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. Virus Evol. 2020;6:veaa027. https://doi.org/10.1093/ve/veaa027
17. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med. 2020;382:2302–15. https://doi.org/10.1056/NEJMoa2006100
18. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill. 2020;25:2000045. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045
19. Quick J. nCoV-2019 sequencing protocol version 1. protocols.io; 2020 Jan 22 [cited 2020 Mar 2]. https://www.protocols.io/view/ncov-2019-sequencing-protocol-bbmuik6w
20. Wick R. Porechop version 0.2.4. 2018 Oct 19 [cited 2020 Jun 18]. https://github.com/rrwick/Porechop
21. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–95. https://doi.org/10.1093/bioinformatics/btp698
22. Simpson J. Nanopolish: signal-level algorithms for MinION data [cited 2020 May 10]. https://github com/jts/nanopolish
23. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90. https://doi.org/10.1093/bioinformatics/bty560
24. Bushnell B. BBMap short read aligner, and other bioinformatic tools [cited 2020 May 10]. https://escholarship.org/uc/item/1h3515gn
25. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34:4121–3. https://doi.org/10.1093/bioinformatics/bty407

26. Rambaut A, Holmes EC, Hill V, O'Toole Á, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. Nat Microbiol. 2020 Jul 15 [Epub ahead of print]. https://doi.org/10.1038/s41564-020-0770-5
27. O'Toole A, McCrone JT. Pangolin: Phylogenetic Assignment of Named Global Outbreak LINeages [cited 2020 Jun 18]. https://github.com/hCoV-2019/pangolin
28. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66. https://doi.org/10.1093/nar/gkf436
29. De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. Issues with SARS-CoV-2 sequencing data. 2020 Jul 29 [cited 2020 Jun 16]. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473
30. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22:160–74. https://doi.org/10.1007/BF02101694
31. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015
32. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21. https://doi.org/10.1093/sysbio/syq010
33. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4:vex042. https://doi.org/10.1093/ve/vex042
34. To T-H, Jung M, Lycett S, Gascuel O. Fast dating using least-squares criteria and algorithms. Syst Biol. 2016;65:82–97. https://doi.org/10.1093/sysbio/syv068
35. World Health Organization. Novel coronavirus (2019-nCoV) situation report—57. 2020 March 17 [cited 2020 May 10]. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
36. United Nations Population Division. World population prospects 2019 [cited 2020 Jun 19]. https://population.un.org/wpp/Download/Metadata/Documentation
37. Candido DDS, Watts A, Abade L, Kraemer MUG, Pybus OG, Croda J, et al. Routes for COVID-19 importation in Brazil. J Travel Med. 2020;27:taaa042. https://doi.org/10.1093/jtm/taaa042
38. Ministry of Health and Social Protection, Colombia. Resolution number 0000380 [in Spanish]. 2020 Mar 10 [cited 2020 June 2] https://www.minsalud.gov.co/Normatividad_Nuevo/Resoluci%C3%B3n%20No.%20380%20de%202020.pdf
39. Ministry of Health and Social Protection, Colombia. Resolution number 0000385 [in Spanish]. 2020 Mar 12 [cited 2020 June 2] https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-385-de-2020.pdf
40. Ministry of the Interior, Colombia. Decree 412 [in Spanish]. 2020 Mar 16 [cited 2020 June 2]. https://dapre.presidencia.gov.co/normativa/normativa/DECRETO%20412%20DEL%2016%20DE%20MARZO%20DE%202020.pdf
41. Ministry of the Interior, Colombia. Decree 457 [in Spanish]. 2020 Mar 22 [cited 2020 June 2] https://dapre.presidencia.gov.co/normativa/normativa/DECRETO%20457%20DEL%2022%20DE%20MARZO%20DE%202020.pdf
42. Ministry of Transport, Colombia. Decree 439 [in Spanish]. 2020 Mar 20 [cited 2020 May 10] https://dapre.presidencia.gov.co/normativa/normativa/DECRETO%20439%20DEL%2020%20DE%20MARZO%20DE%202020.pdf
43. Chancellery of Colombia. Communication on humanitarian flights from 2 to 12 July; 2020 June 17 [cited 2020 Jun 19]. https://www.cancilleria.gov.co/newsroom/publiques/comunicado-vuelos-caracter-humanitario-2-12-julio
44. Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowd-sourced data: a population-level observational study. Lancet Digit Health. 2020;2:e201–8. https://doi.org/10.1016/s2589-7500(20)30026-1
45. Castillo AE, Parra B, Tapia P, Acevedo A, Lagos J, Andrade W, et al. Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. J Med Virol. 2020;92:1562–6. https://doi.org/10.1002/jmv.25797
46. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al.; Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020;369:1255–60. https://doi.org/10.1126/science.abd2161
47. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al.; Sheffield COVID-19 Genomics Group. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;182:812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043
48. da Silva Candido D, Watts A, Abade L, Kraemer MUG, Pybus OG, Croda J, et al. Routes for COVID-19 importation in Brazil. J Travel Med. 2020;27:taaa042. https://doi.org/10.1093/jtm/taaa042

Address for correspondence: Katherine Laiton-Donato; Instituto Nacional de Salud, Sequencing and Genomics Unit, Avenida calle 26 No 51-20, Bogotá DC 110221, Colombia; email: kdlaitond@unal.edu.co, klaiton@ins.gov.co

Article DOI: https://doi.org/10.3201/eid2612.2969

Genomic Epidemiology of Severe Acute Respiratory Syndrome Coronavirus 2, Colombia

Appendix 2

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) by using the incidence i_O for the country o and the total number 〖$P_(o, d)$〗_ of passengers on route

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

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

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

Epidemiologic Investigation Per Lineage

Lineage A.1.2 was detected in the city of Manizales from an imported case in a person 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 3 persons in the city of Medellin without travel history, the first person had symptoms beginning on March 9. These sequences shared a distinctive substitution pattern at the amino acid (aa) G238C at the nucleocapsid and nucleotide, C17470T at ORF1ab helicase, levels (Appendix 1 Tables 2 and 3). A fourth unrelated case from this lineage was identified in the city of Ibague from an imported case in a person arriving on March 17 from an unknown country. Two independent introductions of SARS-CoV-2 lineage A.5 could explain their current epidemiology in Colombia.

We assigned some SARS-CoV-2 sequences to lineage B was assigned because the genetic variability did not enable assignment to a specific sublineage. One patient, Colombia/Cali/79449, arrived in the city of Cali from Spain on March 7. The viral sequence displayed a very similar pattern to lineage B.2, with 2 aa changes, L3606F at open reading frame 1ab (ORF1ab; Nsp6) and G251V at ORF3a (Appendix 1 Table 3). The other 2 patients were identified in the same urban area, Medellín and Sabaneta municipalities belong to the Aburra Valley metropolitan area; the first of the 2 had symptoms beginning on March 13. These results suggest 2 independent introduction events to explain the current presence and distribution of SARS-CoV-2 basal lineage B in Colombia.

Among SARS-CoV-2 sequences analyzed from Colombia, lineage B.1 represented 48%. The genomic analysis of the genetic variability at the aa level revealed 4 substitution patterns (Appendix 1 Table 3), which increased to 5 when analyzed at the nucleotide level (Appendix 1 Table 2). The first pattern was identified in 2 sequences from the city of Neiva belonging to the same transmission chain without travel history. The second substitution pattern was identified mainly in 8 municipalities from the Valle del Cauca department and most sequences were collected from the capital city, Cali. These sequences displayed a distinctive aa change in T265I at ORF1ab (Nsp2) (Appendix 1 Table 3). Eight patients reported international travel history from Spain and the United States and the earliest arrived on March 6, 2020. The third substitution

pattern mainly was distributed in the Caribbean region, in the cities of Ciénaga, Monteria, Santa Marta, and Valledupar, and municipalities from the center of the country near the capital, Bogotá. The first case belonging to this pattern corresponded to an imported case in a person entering the country from Spain on March 1, early during the pandemic in Colombia and previous to the domestic and international flights restriction. The fourth substitution pattern was identified in several municipalities without geographic proximity, including Medellín, Cartagena, Ibague, Leticia, and Togui (this case was reported in a person with recent travel to a national tourist region), from confirmed cases without international travel history from the available data. The first case among this group of sequences was identified in Cartagena on March 10, 2020. Finally, the fifth substitution pattern was identified in 3 sequences from samples collected in the cities Neiva and Envigado, which are in different regions of the country. The first patient reported symptoms that began on March 18, 2 days after the person entered Colombia from Panama. These sequences displayed a distinctive aa change in I71V at ORF1ab (leader). Patterns 2–5 shared the aa change Q57H at ORF3a; patterns 3–5 shared the synonymous substitution C18877T, 3′-to-5′ exonuclease; and patterns 4 and 5 shared the synonymous substitution C10509T, Nsp5A (Appendix 1 Tables 2 and 3).

Lineage B.1.1 was not directly assigned by PANGOLIN (https://github.com/cov-lineages/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 aa changes R203K and G204R (Appendix 1 Table 3; 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 in a person entering the country from Italy on February 26 and displays 5 nucleotide substitutions common to pairs of sequences, 4 of which also displayed geographic correspondence; thus, representing previously identified or unidentified local transmission chains. Two sequences obtained from patients from Villavicencio city without traveling history belonged to a previously characterized transmission chain and displayed 2 distinctive aa changes, S2488F at Orf1ab (Nsp3) and T14A at ORF7a (Appendix 1 Table 3).

Lineage B.1.3 was identified in the city of Pereira and the municipality of Dosquebradas, which belong to the same metropolitan area. The patients had no traveling history and the first one had symptoms beginning on April 2. Despite the lack of available information about the

epidemiologic relationship between the 2 patients, the fact their sequences shared the same nucleotide substitution pattern, geographic distribution, and temporality suggest that they belong to the same transmission cluster.

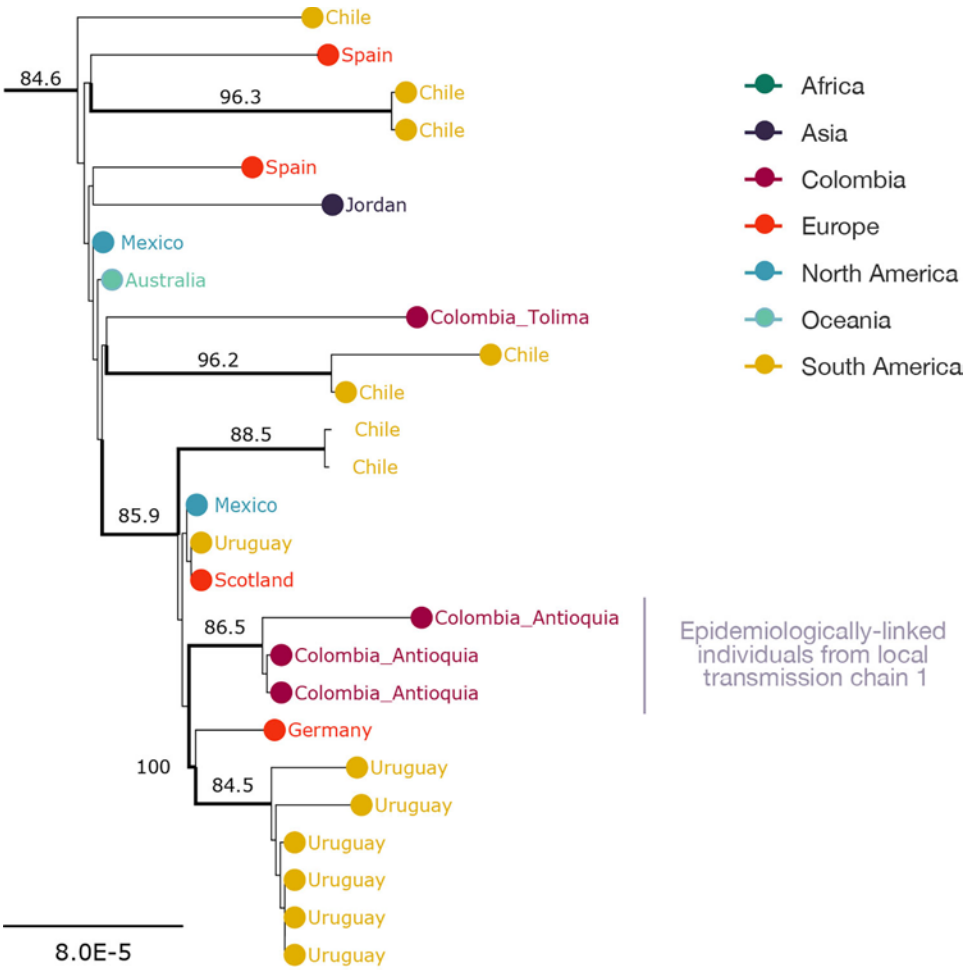
Lineage B.1.5 was the second most frequently noted sequences, accounting for the 26% of the analyzed sequences from Colombia. Any aa change is distinctive of this lineage, but all the sequences shared the exclusive synonymous substitution of A20268G at ORF1ab (endoRNase). The analysis of the genetic variability at the amino acid level revealed 3 substitution patterns (Appendix 1 Table 3), which increased to 5 when analyzed at the nucleotide level (Appendix 1 Table 2). The first substitution pattern was represented by sequences from the municipalities of Barranquilla, Bello, Bogotá, Cali, Cucuta, Medellín, Neiva, Pacho, Palmira, and Tierralta, located in 7 different departments. Despite the lack of geographic clustering, 12/18 patients belonging to this group reported entering the country from Spain, France, or Italy. Of note, 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 in the Spike protein gene. According to the available information from 2 of the patients, no travel history was reported. 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 also reported entering the country from Spain on March 9, 2020 and having symptoms on the same day. The fourth pattern was identified in the city of Cali from 2 patients belonging to the same transmission chain who shared the aa change R191C at the Nucleocapsid protein and the synonymous substitution C1327T at Orf1ab (Nsp2). The fifth substitution pattern was defined by two distinctive aa changes at Orf1ab, at A3610V in the Nsp6 and at G5063S in RdRp genes. This pattern was identified in Cali city and Buga municipality, both in the department of Valle del Cauca, the first patient reported travel history from Spain on March 13, 2020.

Lineage B.1.8 was identified in the city of Pereira 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 who reported symptoms starting on March 23. Lineage B.2 was identified in the city of Cali from a single sequence obtained from a patient without travel history who reported symptoms beginning on March 29. Finally, lineage B.2.5 was

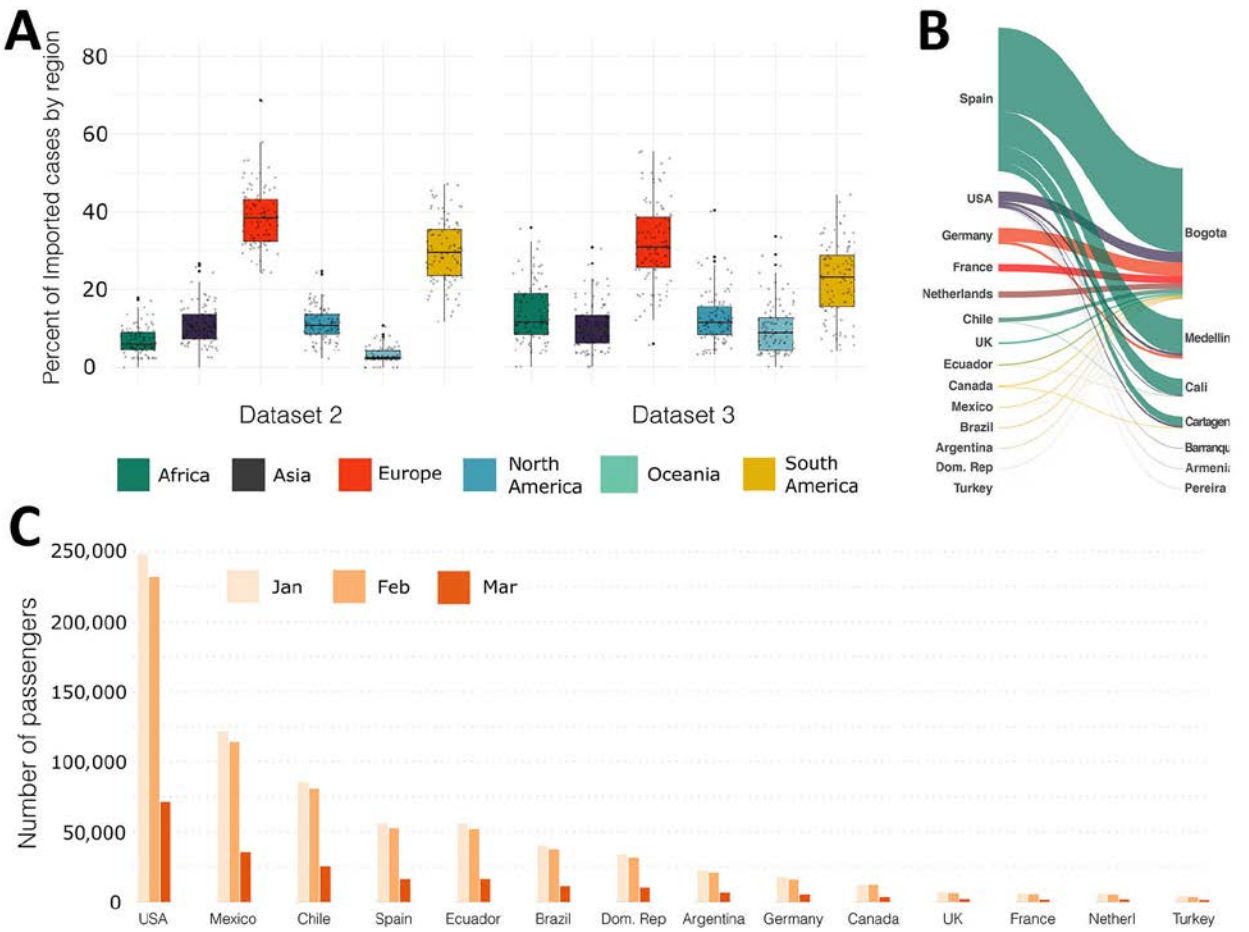
represented by a previously referred transmission chain with the first patient arriving in Armenia on March 10, from Italy.

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 but estimates for within each lineage sequence pairs were 0.000169 for lineage A and 0.000222 for lineage B (Appendix 1 Table 4). 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 sublineage level 1 enable sequences to be grouped into A.x and B.x and at sublineage level 2, the grouped as belonging to A.x.x and B.x.x. These comparisons at sublineage level 1 enabled us 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. Of note, the average divergence between sublineages assigned to the same lineage, such as (A.1 vs. A.2 or A.2 vs. A.5, was higher than other comparisons between sublineages assigned to different lineages, such as A.1 vs. B.1, A.1 vs. B Basal, or A.2 vs. B Basal. The higher resolution comparisons at sublineage level 2 enabled us 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 with any sublineage from the B lineage and also when B.2.5 was compared with any other sublineage. The average divergence between sublineages assigned to the same lineage, including A.1.2 vs. A.2 Basal, A.2 Basal vs. A.5 Basal, B.1.1 vs. B.2.5, and B.1.3 vs. B.2.5, was higher than other comparisons between sublineages assigned to different lineages, such as A.1.2 vs. B.1.5, B.1.8, B.1.11, B.1 Basal, B.2 Basal, or B Basal (Appendix 1 Table 4). These results can be affected by the low sequence number for some lineages, but the estimates of the evolutionary distance between groups are in agreement with the lineage assignment and with the substitution patterns observed to be unique to each lineage (Appendix 1 Tables 3, 4).



Appendix 2 Figure 1. Example of a SARS-CoV-2 clade with sequences from a local transmission chain from Colombia. Tips are labeled by country (and department in the case of isolates from Colombia) and colored by region. Of note, fine-grained relationships could not always be resolved with confidence. Branch supports with aLRT <0.75 are not shown. Scale bars indicate number of nucleotide substitutions per site. aLRT, approximate likelihood ratio test; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.



Appendix 2 Figure 2. Potential countries of introductions of SARS-CoV-2 to Colombia. A) Geographic source attribution for every transition into Colombia derived from the migration inference retaining several sequences per region, when possible, (dataset 2) equal to the number of sequences available for Colombia and retaining 50 sequences per region and all sequences from Colombia (dataset 3). Error bars show SDs; horizontal lines indicate means. B) Number of passengers arriving in Colombia from different countries. C) Geographic contribution inferred by using air travel data per country.