Received: 2 June 2021 | Revised: 23 June 2021 | Accepted: 25 June 2021

DOI: 10.1002/jmv.27167

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Phylogenomics reveals multiple introductions and early spread of SARS-CoV-2 into Peru

Eduardo Juscamayta-López[1,2] | Dennis Carhuaricra[2] | David Tarazona[1,2] | Faviola Valdivia[1] | Nancy Rojas[3] | Lenin Maturrano[2] | Ronnie Gavilán[4,5]

[1]Laboratorio de Infecciones Respiratorias Agudas, Centro Nacional de Salud Pública, Instituto Nacional de Salud, Lima, Peru

[2]Laboratorio de Biología y Genética Molecular, Universidad Nacional Mayor de San Marcos, Lima, Peru

[3]Laboratorio de Virus Respiratorios, Centro Nacional de Salud Pública, Instituto Nacional de Salud, Lima, Peru

[4]Laboratorio de Enteropatógenos, Centro Nacional de Salud Pública, Instituto Nacional de Salud, Lima, Peru

[5]Escuela Profesional de Medicina Humana, Universidad Privada San Juan Bautista, Lima, Peru

Correspondence

Eduardo Juscamayta-López, Laboratorio de Infecciones Respiratorias Agudas, Centro Nacional de Salud Pública, Instituto Nacional de Salud, Cápac Yupanqui 1400 - Jesús María, Lima 11, Peru.
Email: jjuscamayta@ins.gob.pe and ejuscamaytal@gmail.com

Funding information

Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica; National Institute of Health of Peru and the CONCYTEC-FONDECYT Program of Proyectos Especiales: Respuesta al COVID-19 2020-01-01: 034-2020-FONDECYT

Abstract

Peru has become one of the countries with the highest mortality rates from the current coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). To investigate early transmission events and the genomic diversity of SARS-CoV-2 isolates circulating in Peru in the early COVID-19 pandemic, we analyzed 3472 viral genomes, of which 149 were from Peru. Phylogenomic analysis revealed multiple and independent introductions of the virus likely from Europe and Asia and a high diversity of genetic lineages circulating in Peru. In addition, we found evidence for community-driven transmission of SARS-CoV-2 as suggested by clusters of related viruses found in patients living in different regions of Peru.

KEYWORDS

evolution, molecular epidemiology, phylogenomics, SARS-CoV-2, transmission, whole-genome sequencing

1 | INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV 2) is a novel Betacoronavirus responsible for coronavirus disease 2019 (COVID-19) that originated in December 2019 in Wuhan, China.[1] Since then, the virus has rapidly spread across the globe and was declared a pandemic on March 11, 2020.[2] Immediately, the Peruvian government carried out a series of sanitary interventions to prevent disease transmission including schools' closure (March 11, 2020) and travel restrictions from Europe and Asia (March 12, 2020), country border closures (March 16, 2020), and a nationwide curfew (March 18, 2020).[3] Despite these measures, SARS-CoV-2 has managed to rapidly spread causing over 702,776 cases and 30,236 deaths with Lima city as one of the major epicenters of SARS-CoV 2 infections in Peru, as of September 11, 2020.[4] As a consequence, Peru has become one of the countries with the highest mortality rate from the current COVID-19 pandemic. SARS-CoV-2 whole-genome sequence (WGS) data has rapidly become publicly available, revealing insights into genome structure as well as the temporal evolution and global transmission of the virus.[5] However, the sources of epidemic transmission and genomic diversity of SARS-CoV-2 strains circulating in Peru in the early COVID-19 pandemic remains poorly investigated.

We analyzed a total of 3472 SARS-CoV-2 genomes (149 from Peru) to investigate how this novel virus became established in the country and to dissect its spread in this area, which will help to orient effective prevention measures to control the COVID-19 epidemic in Peru.

2 | MATERIALS AND METHODS

2.1 | Sample collection and WGS of SARS-CoV-2

We obtained RNA from nasal and pharyngeal swab samples collected from confirmed COVID-19 patients at the National Institute of Health, Peru (NIH-Peru), between March 5, 2020, and July 4, 2020, from departments of Lima, Callao, Ancash, Lambayeque, Ica, Arequipa, and Junin. Cases were randomly selected regardless of disease severity and hospital origin from samples that tested positive for SARS-CoV-2 RNA by quantitative reverse-transcription polymerase chain reaction and with Ct less than 25. The WGS of SARS-CoV-2 isolates ($n = 96$) were performed on MiSeq (Illumina) at NIH-Peru using the CleanPlex® SARS-CoV-2 Panel (Paragon Genomics) through amplicon-based target enrichment.

2.2 | Sequence quality analysis and SARS-CoV-2 genome assembly

The quality of the obtained sequences was evaluated through the software FastQC v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Contaminant residues, low-quality, and duplicated reads were trimmed off using the program fqCleaner v.0.5.0, with a Phred quality score of 28. Filtered reads were mapped against SARS-CoV-2 reference (NC_045512) using the Burrows-Wheeler Aligner MEM algorithm (BWA-MEM) v0.7.7. SAMtools v1.9[6] and Geneious Prime were used to sort BAM files, generate alignment statistics, and obtain consensus sequences. Previous to the construction of the consensus sequence, we removed primer adapters with the software fgbio (https://github.com/fulcrumgenomics/fgbio) using the primer sequence coordinates provided by Paragon Genomics.

2.3 | Peruvian and global collection of SARS-COV-2 genome sequences retrieved from GISAID

The SARS-CoV-2 genomes from Peru ($n = 62$) available in GISAID as of August 26, 2020, were selected getting a data set of 149 Peruvian genomes that cover a period from March 5, 2020 (officially first reported case) to July 4, 2020, of 12 Peruvian regions. To place the Peruvian sequences into a global context and avoiding dealing with a large data set computationally intractable, we retrieved sequences from other countries by means of a subsample ($n = 3323$) from viral genomes publicly available in GISAID randomly choosing one genome per country per day (collection date) between December 24, 2019 (including Wuhan-01 genome) and July 1, 2020. The final genomic data set comprised 3472 sequences to investigate the origins and genomic diversity of SARS-CoV-2 isolates circulating in Peru by Maximum-Likelihood (ML) phylodynamic approaches.

2.4 | Phylogenetic analysis of SARS-CoV-2 in Peru

The full genomic data set ($n = 3472$) was aligned using MAFFT v7.1[7] with default parameters. The alignment was manually curated, trimming the 5′ and 3′ ends and ambiguous regions obtaining an alignment length of 29520 bp. We estimate an ML tree of 3472 aligned sequences using IQ-tree v1.6[8] under the HKY nucleotide substitution model, with a gamma distribution of among-sites rate variation (HKY + G + I) as selected by ModelFinder.[9] The EPI_ISL_406801 sequence (Wuhan/WH04/2020), a basal A lineage sequence, was used as an outgroup for the ML tree.[10] To measure branch support, we used the Shimodaira–Hasegawa and approximate likelihood-ratio test with 1000 replicates. TempEst v1.5.1[11] was used to assess the strength of the temporal signal and inspect for outliers in the data set by a root-to-tip regression of genetic distance against sampling date. We used the program TreeTime v0.7.6[12] to estimate an ML time tree. The analysis was performed using the HKY substitution model and a coalescent Skyline prior under a strict clock.[13] We also passed the flag–confidence to retrieve node dates with 90% confidence intervals (CIs).

The clades were analyzed with Next Strain classification for SARS-CoV-2 (https://nextstrain.org/ncov) and clade-specific nonsynonymous mutations were identified. Introductions or local transmission events were designated as nodes (1–18) with more than or equal to 70% of statistical support value in the ML tree and the time to the most recent common ancestor (tMRCA) was estimated for each node with 90% CIs. Introduction events were defined as Peru sequences that clustered with non-Peru sequences across different clades. Local transmission events were defined as Peru-exclusive clusters with at least five or more sequences that were reproduced in the time-scaled inference.[14] Statistical analyses and tree visualization using ggtree package were carried out in R v3.6.2.

The lineages were determined using the nomenclature of Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN) (https://pangolin.cog-.io/).[15] Next strain classification helps to reference origins of sequence patterns while the nomenclature of PANGOLIN provides a convenient scheme for genomically detectable introductions of SARS-CoV-2 into new regions.

3 | RESULTS AND DISCUSSION

Peru has one of the world's highest levels of contagion and deaths by the COVID-19 pandemic despite early national lockdown decreed by the government on March 15, 2020, when the country just had 71 cases reported. We sequenced 96 SARS-CoV-2 genomes obtained from patients with confirmed COVID-19 diagnosis up to July 4, 2020, at NIH-Peru, yielding 87 high coverage-quality genomes. We removed nine sequences with low coverage or too many private mutations/ambiguous sites indicative of sequencing error. These 87 cases were drawn from seven Peruvian regions including Lima, Callao, Ancash, Lambayeque, Ica, Arequipa, and Junin. Additionally to these sequences, we obtained 62 Peruvian SARS-CoV-2 genomes

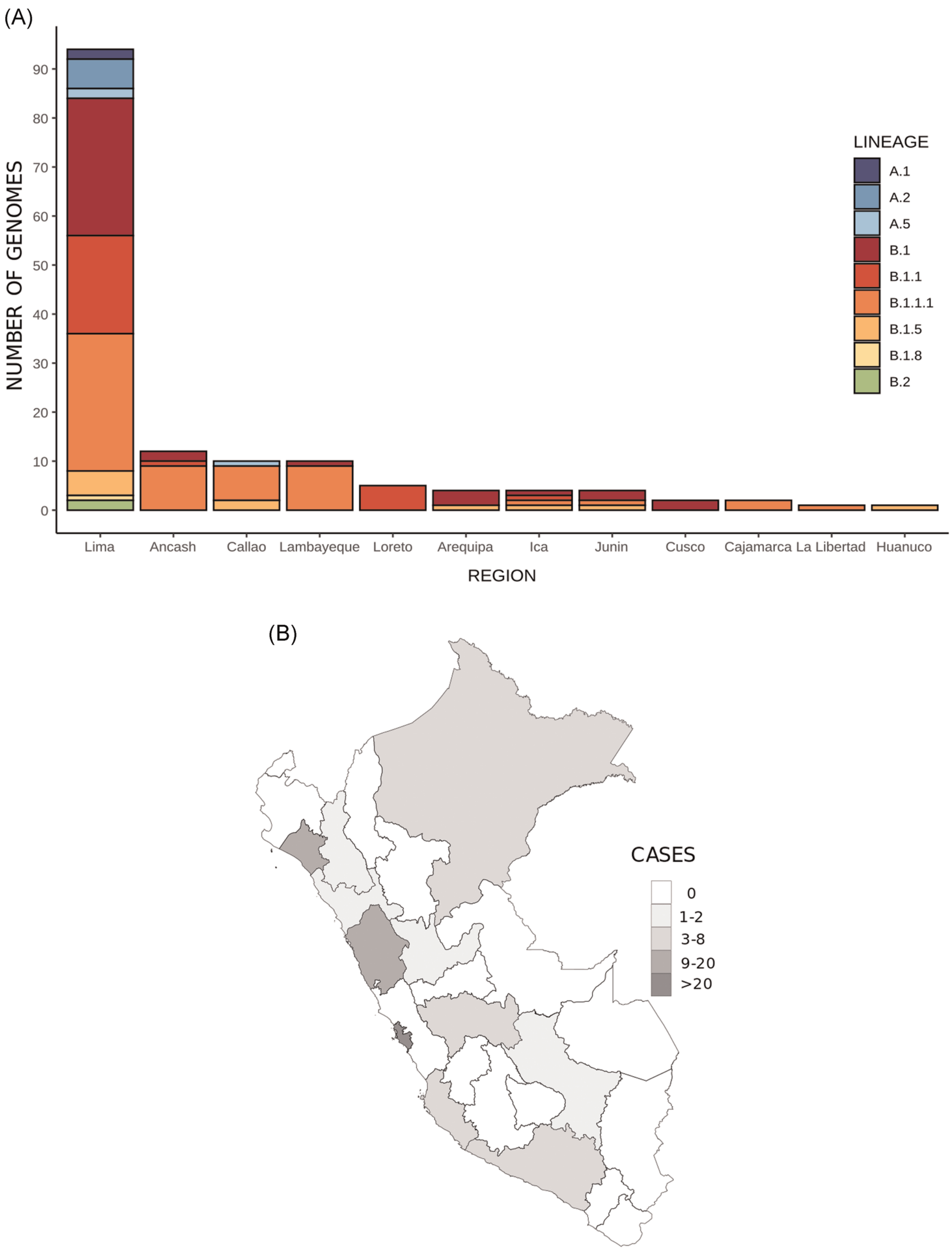

FIGURE 1 Geographic distribution of the Peruvian SARS-CoV-2 isolates circulating across regions of Peru. (A) Distribution of sequenced cases across Peru departments. (B) Breakdown of sequenced cases according to phylogenetic clades across Peru departments. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2

from the GISAID database, bringing our total genomic data set to 149 Peruvian isolates from 12 regions that cover a period from the first official case on March 5, 2020, until July 4, 2020 (Figure 1A). Total Peruvian isolates were obtained from nasal and pharyngeal swab samples collected from 71 females (47.65%) and 78 (52.35%) males ranging in age between less than 25 years (14.37%), 25–35 years (17.36%), 36–46 years (22.65%), 47–59 years (21.56%), and more than 60 (24.06%) years.

Ninety-four (63%) out of 149 Peruvian sequences were obtained from Lima Metropolitana, the Peruvian capital, and 10 genomes from Callao, a neighbor city of Lima, where the Jorge Chavez International Airport is located and the only one in the country with flights to and from Asia, Europe, and North America (Figure 1A). Moreover, Lima-Callao is considered the epicenter of the COVID-19 pandemic for Peru presenting 47% of total cases until September 5, 2020.[4] Our genomic data set also contains 25 isolates from regions located in the North of Peru (Lambayeque, La Libertad, Ancash, and Cajamarca), 10 sequences from southern regions (Ica, Arequipa, and Cuzco), five genomes from Loreto (the hardest-hit region by the pandemic in the Peruvian Amazon), and five from central regions of Peru (Junin and Huanuco) (Figure 1B).

We constructed a time-scaled phylogenetic tree from 3472 SARS-CoV-2 genomes sampled between December 24, 2019, and July 4, 2020, including 149 Peruvian isolates (Figure 2A) to investigate early transmission events and genomic diversity of SARS-CoV-2 strains circulating in Peru. We found a strong temporal structure in our data set by regressing genetic distance from tips to root in the ML tree against sample dates, resulting in a relatively high positive correlation ($r^2 = 0.51$) (Figure S1). ML phylodynamic analysis revealed that the tMRCA of the analyzed full genomic data was November 28, 2019 (November 20, 2019–December 14, 2020, 90% CI) and the inferred ancestral root being Asia. These observations are in line with the known epidemiology of the pandemic.[5] Furthermore, the analysis of genomic data has shown that Peruvian isolates were widely distributed across the phylogenetic tree suggesting multiple and independent introductions, designed as nodes (1–18) with over 70% of statistical support from ML (Figure 2A and Table 1). Similar introductions have been reported in other countries such as Brazil,[16] Colombia,[17] and the USA.[14] Most of these putative introductions of SARS-CoV-2 into Peru occurred between mid-February and early March, likely sourced from Europe, Asia, North America, and South America (Table 1). Concordantly, the COVID-19 pandemic reached Latin America in February 2020 expanding into the region until March 2020, when the COVID-19 incidence curve started to grow more rapidly.[3] Also, our results coincide with initial analysis based on minimum spanning tree (MST) of 36 Peruvian SARS-CoV-2 genomes that suggested the introduction of multiple isolates from Europe and Asia.[18] However, similar to genome-based analyses, the MST model must be supported by an exhaustive clinical-epidemiological investigation to identify the root of transmission in an outbreak of infectious diseases.[19]

Peruvian isolates (57%) were mainly clustered with clade 20B according to the Next strain nomenclature. This clade is predominantly composed of isolates obtained from patients with COVID-19 in Europe (51%), suggesting that introductions from Europe account for the majority of cases found in Peru between February and early March 2020 (Figure 2B and Table 1). Surprisingly, as of March 12, 2020, the Ministry of Health reported 11 confirmed cases (out of 22 total cases) in Lima with recent travel history to Spain, France, and Italy.[20]

We also identified within clade 20B, five putative introductions (node 1–5) where the most Peruvian isolates were interspersed without grouping by country or geographic regions. Within this clade, we also identified SARS-CoV-2 Peruvian sequences that were reproduced in the time-scaled inference with mutations T1246I and G3278S in the ORF1a gene that distinguish clusters of sequences from Peru and elsewhere, suggesting a local transmission event that likely occurred in early March (tMRCA 90% CI: 27 February–2 March) (Table 1, node 1, Figures 2A and S1B).

Similar to clade 20B, SARS-CoV-2 isolates positioned in clade 20A were distributed among isolates from multiple regions and mainly composed of European isolates (54%) (Figure 2 and Table 1). These results are in line with the earliest sequences observed at the base of this clade (France, Russia, and the Czech Republic). Thus, it is highly likely to be sourced of European origin. Within this clade, we also identified a Peruvian cluster (Table 1, node 10) which includes the first official COVID-19 case in Peru (Peru/INS-01–02) identified in Lima City on March 5, 2020, and with travel history to France, Spain, and the Czech Republic.[20] The SARS-CoV-2 sequence of this patient clustered with other Peruvian isolates (from March 6, 2020, to July 4, 2020) conforming a "Peruvian cluster," all of which were closely related with European isolates confirming its likely origin (Figures 2 and S1B).

The tMRCA of this cluster (node 10) was estimated to be March 4, 2020 (February 25, 2020–March 4, 2020, 90% CI) and contains two specific amino acid substitutions: H604Y in the ORF1b gene and I119V in the spike protein (Table 1). Within the cluster, we also observed 6 identical genomes to Peru/INS-01-02 reported in Lima until March 17, 2020, suggesting local spread (Figures 2C and S1B). Interestingly, two identical sequences to Peru/INS-01-02 reported in Cuzco, in southern Peru, on March 10, 2020, were identified, suggesting a regional introduction of SARS-CoV-2 derived from the first case from Lima (Figures 2C and S1B). Although by March 18, 2020, the Peruvian government had implemented a nationwide curfew, the phylogenomic analysis revealed cases derived from the "Peruvian cluster" reported between April and July to Lima, Ica, and Junin, suggesting an epidemic spread of SARS-CoV-2 into Peru's regions likely by local mobility.[21] These results, together with the nationwide widely distributed Peruvian isolates within a clade, support that the community spread likely originated from Lima. However, it should be noted that although genetic linkage provides evidence of community transmission, epidemiological linkage is required for confirming it.

For the remaining clades (19A, 19B, and 20C), we inferred SARS-CoV-2 virus putative introductions to Peru between mid-December 2019 and early March 2020 (Figure 2A and Table 1). The clades 19A and 19B have mainly included Asian isolates (72% and 54%, respectively) suggesting that likely introductions from Asia account for the majority of

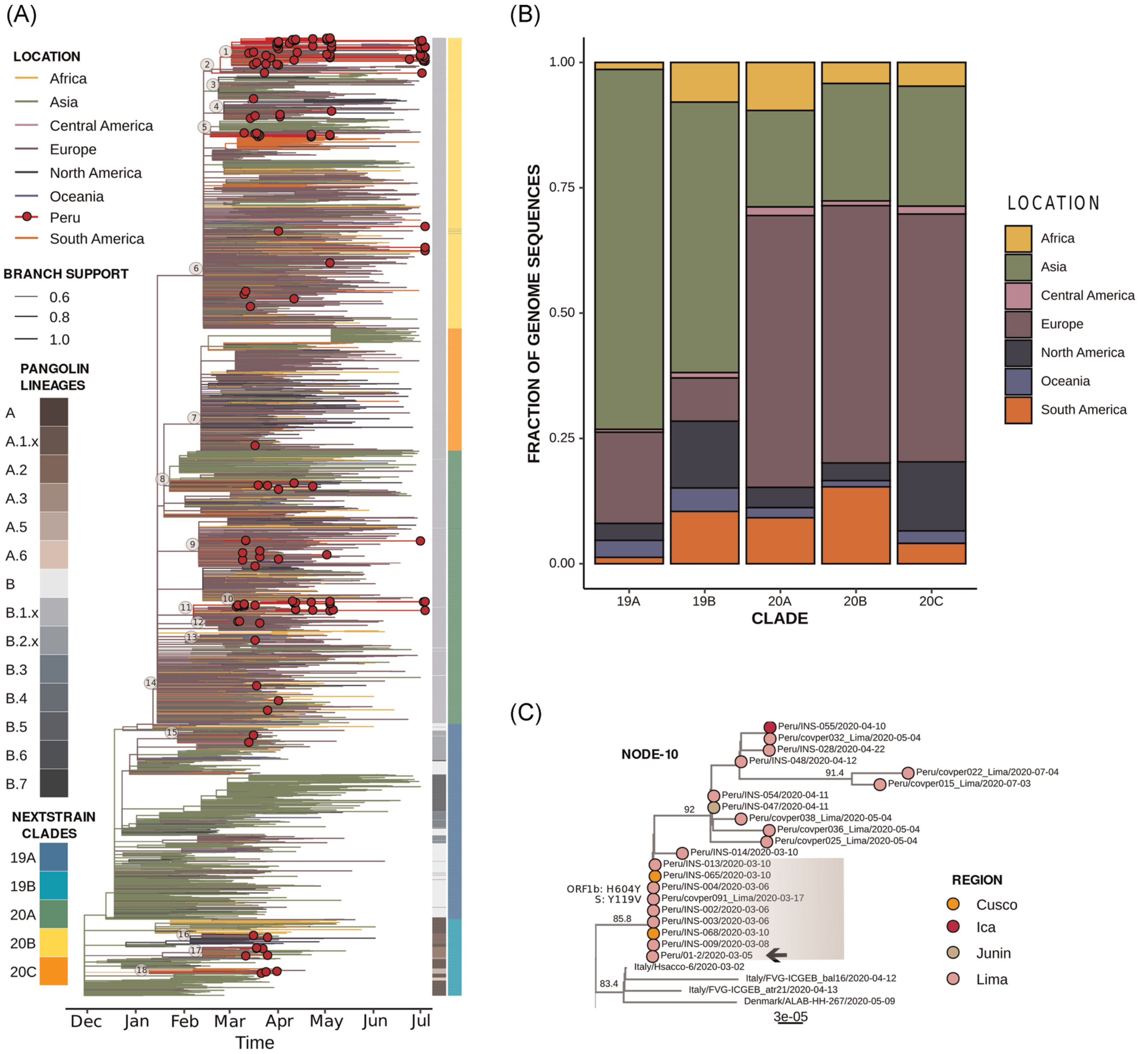

FIGURE 2 Phylogenetic relationships of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) from Peru and other global strains. (A) Maximum-Likelihood phylodynamic inference of 149 SARS-CoV-2 sequences from Peru in a global background of 3323 sequences available in the GISAID database as of 4 July 2020. Branches are colored according to the region of origin. Tip circles (red) indicate the position of the 149 Peruvian isolates. Clades that contain Peru sequences are highlighted, with names shown on the left. The node positions for the transmission events are marked by the numbers in white circles. (B) Stacked bar plot showing the fraction of genome sequences per location by clade.
(C) Local transmission clusters on the ML tree showing the source of cases by departments. Bootstrap support values ≥ 70% are shown; sibling clusters are collapsed for easier visualization. The mutations identified specific to the community transmission cluster are indicated (ORF1b: H604Y; S: Y119V). The scale bar at the bottom indicates the number of nucleotide substitutions per site

cases found in Peru in the early pandemic period (Figure 2A,B and Table 1).

To identify the genetic diversity of SARS-CoV-2 circulating in Peru, all Peruvian sequences were classified using the PANGOLIN nomenclature resulting in nine lineages: A.1, A.2, A.5, B.1, B.1.1, B.1.1.1, B.1.5, B.1.8, and B.2. The main lineages found in Peru were B.1 and B.1.1.1 grouping 97 (65%) of all 149 Peruvian sequences. Similar results were found by Padilla et al. in a study based on 39 Peruvian sequences identifying a relatively high prevalence of the viral lineage B.1 (38%, $n = 13$), and the viral sub-lineage B.1.1.1 (29.4%, $n = 10$).[18] Previous studies have reported relevant differences in the geographical distribution of the three most represented lineages (B.1, B.1.1, and B.1.5) in Lombardy, Italy.[22] Conversely, in our study, Lima-Callao presents all nine lineages found in Peru, whereas the rest of the region's harbor lineages B.1, B.1.1, B.1.1.1, and B.1.5 (Figure 1B). Additionally, until March 2020, we observed the nine PANGOLIN lineages in Lima-Callao including A lineages; however, after April 2020, only B.1.1.1, B.1.1, B.1, and B.5 were registered by WGS (Figure 2A).

TABLE 1 Inferred SARS-CoV-2 virus transmission events related to Peru

Next strain clade	First Peru sequence	Nonsynonymous mutations	Node	Cases	Event type	Inferred putative source	tMRCA 90% CI
20B	03-16-20	ORF1a: C4002T[T1246I], G10097A[G3278S]	1	56	Putative introduction	Europe	Feb 27–Mar 2
	03-23-20	None	2	2	Putative introduction	Europe	Feb 14–Feb 18
	03-16-20	None	3	1	Putative introduction	Europe/Asia	Feb 17–Feb 23
	03-14-20	None	4	5	Putative introduction	Europe	Feb 21–Feb 25
	03-10-20	ORF1a: G10265A [G3334S]	5	12	Putative introduction	Europe	Feb 13–Feb 24
	03-10-20	None	6	9	Untracked transmission	Europe	Feb 9–Feb 14
20C	03-17-20	ORF1a: C1059T[T265I] ORF3: G25563T[Q57H]	7	1	Putative introduction	Europe	Feb 7–Feb 15
20A	03-19-20	ORF3a: G25563T[Q57H]	8	6	Putative introduction	Europe	Jan 19–Jan 22
	03-09-20	None	9	10	Putative introduction	Europe	Feb 4–Feb 21
	03-05-20	ORF1b: C15277T[H604Y] S: A21917G[I119V]	10	21	Putative introduction	Europe	Feb 25–Mar 4
	04-12-20	ORF1b: C18568T[L1701F]	11	6	Putative introduction	Europe	Mar 8–Apr 12
	03-06-20	ORF3a: G25429T[V13L]	12	3	Putative introduction	Europe	Jan 31–Feb 29
	03-17-20	None	13	1	Putative introduction	Europe	Jan 21–Feb 27
	03-18-20	None	14	3	Untracked transmission	Europe	Jan 9–Jan 14
19A	03-13-20	None	15	2	Putative introduction	Asia/Europe	Jan 15–Feb 6
19B	03-16-20	ORF1b: C17747T[P1427L], A17858G[Y1464C]	16	2	Putative introduction	Asia/North America	Feb 10–Mar 1
	03-11-20	ORF1: T9477A[F3071Y]; ORF3a: G25979T[G196V]; ORF8: T28144C [L84S]; N: C28863T[S197L], ORF14: Q44*	17	6	Putative introduction	Asia/Europe	Feb 5–Feb 18
	03-21-20	None	18	3	Putative introduction	Asia/South America	Dec 18, 2019–Jan 31, 2020

Abbreviations: CI, confidence interval; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; tMRCA, time to the most recent common ancestor.

A similar behavior pattern had the lineages of SARS-CoV-2 in Chile starting with the lineage A.5 in early March, turning to the predominant sub-lineages B.1 (41.8%), B.1.1 (20.5%), and B.1.5 (18.4%) in April.[23]

Furthermore, we identified non-synonymous mutations in all Peruvian SARS-CoV-2 sequences from sampled regions from which the most frequent and relevant mutation was the D614G (*S* gene, $n = 136/149$) (Figure S2). The spread of B sub-lineages into Peru might be associated with the nonsynonymous mutation D614G at spike protein, hypothesizing that the D614G mutation provides the virus to be more infectious[24] as there is evidence that the G614 variant may enhance the virus transmission by increasing viral load in the upper respiratory tract from patients with COVID-19.[25] These results are supported by the global epidemiological information that showed the predominance of the G614G variant after SARS-CoV-2 spread into Europe in late February and March. Concordantly, most introductions into South American countries (e.g., Colombia or Brazil) included European B lineage isolates showing the predominance of the G614 variant in March and April.[16,17,24,26] It is known that the carriage of SARS-CoV-2 mutations into naive populations might be due to the neutral founder effect. Nevertheless, the ratio of nonsynonymous to synonymous mutations in S protein (0.25–0.5) is consistent with an emerging virus undergoing purifying selection.[27] However, any other factors might be involved in the global predominance of the G614 variant including uneven sampling, chance, and epidemiological reasons.[28] Although the G614 variant has not been associated with disease severity, it correlated strongly with the mortality rate of COVID-19 (Pearson's correlation coefficient (r) = 0.43, $p = 0.022$) during the early pandemic in a global survey of countries.[29,30]

One of the limitations of this study is the difference in the sequenced genomes between countries. It is well known that some locations (e.g., UK) are sequencing SARS-CoV-2 genomes much more intensively than others while other countries are not sequencing at all. This means that some locations may appear to be infection sources simply because they provide more sequences, rather than actually being the source of infection for a particular disease. In addition, due to the relatively long incubation period and low mutation rate of SARS-CoV-2, two genome sequences may be identical despite having last shared a common ancestor 4–6 weeks ago—during which time the pathogen might have been transmitted through one or multiple missed nodes and locations (e.g., Germany to the USA to Peru, rather than directly Germany to Peru).

Another limitation is the differences between the number of Peruvian isolates from cases identified within Peru regions. Specifically, there were more SARS-CoV-2 representative' sequences in Lima than in the other Peruvian departments. Moreover, in the absence of epidemiological information such as travel history and contact tracking, it is hard to associate periods of untracked transmissions with any specific regions or countries.

Although our estimates might be biased due to the abovementioned, it was possible to obtain an important group of samples that allowed getting valuable information about the spread and genomic diversity of the virus during the early COVID-19 pandemic in Peru.

4 | CONCLUSION

We provide a first snapshot of the sources of epidemic transmission and genomic diversity of SARS-CoV-2 lineages circulating in Peru during the early COVID-19 pandemic, revealing multiple and independent introductions of the virus likely from Europe and Asia, and high diversity of genetic lineages. In addition, we found evidence that the early spread of the virus in Lima City was sustained by community transmission. The data also underscore the limited efficacy of travel restrictions in a place once multiple introductions of the virus and community-driven transmission had already occurred.

Finally, we highlight the need for early and continuous nationwide testing as well as the implementation of a real-time national surveillance system based on Whole-Genome Sequencing to identify untracked transmission clusters and to unscramble the evolution of the SARS-CoV-2 virus into Peru and assess its impact on disease severity, all of which will provide insights on Public health interventions to limit and control the spread of SARS-CoV-2.

ACKNOWLEDGMENTS

We are greatly grateful to all health personnel of Peru, especially to members of the Respiratory Virus Laboratory from NHI-Peru for their dedication in providing continued diagnostic and high-quality care during the pandemic. The authors would like to express our deep gratitude to Dr. Kelly Levano for his critical review of this manuscript. In addition, we wish to thank Guillermo Trujillo and members of the Respiratory Infectious Diseases Laboratory from NHI-Peru, especially Helen Horna and Liza Linares for their support for the Whole-Genome Sequencing. This work was supported by the National Institute of Health of Peru and the CONCYTEC-FONDECYT Program of Proyectos Especiales: Respuesta al COVID-19 2020-01-01 [grant number 034-2020-FONDECYT].

CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest.

ETHICS STATEMENT

The use of human biological material was approved by the Ethics Committee of the National Institute of Health from Peru (OI-014-20) and all experiments were performed in accordance with the ethical standards of the Declaration of Helsinki or comparable ethical standards.

AUTHORS CONTRIBUTIONS

Eduardo Juscamayta-López conceived the study design and data analysis and wrote the manuscript. Faviola Valdivia, Nancy Rojas, and Ronnie Gavilán helped in the sample processing and data collection. Eduardo Juscamayta-López, Dennis Carhuaricra, David Tarazona, and Lenin Maturrano performed the bioinformatic analysis, data processing, and interpretation. All authors critically reviewed the manuscript. Eduardo Juscamayta-López supervised the study.

ORCID

Eduardo Juscamayta-López http://orcid.org/0000-0001-6843-3206

Faviola Valdivia https://orcid.org/0000-0001-8347-6853

REFERENCES

1. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):265-269. https://doi.org/10.1038/s41586-020-2008-3
2. Li X, Giorgi EE, Marichannegowda MH, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv*. 2020; 6(27):eabb9153. https://doi.org/10.1126/sciadv.abb9153
3. Munayco CV, Tariq A, Rothenberg R, et al. Early transmission dynamics of COVID-19 in a southern hemisphere setting: Lima-Peru: February 29th–March 30th, 2020. *Infect Dis Model*. 2020;5:338-345. https://doi.org/10.1016/j.idm.2020.05.001
4. Ministry of Health. Sala situacional COVID-19 Peru. Published 2020. https://covid19.minsa.gob.pe/sala_situacional.asp
5. Li J, Li Z, Cui X, Wu C. Bayesian phylodynamic inference on the temporal evolution and global transmission of SARS-CoV-2. *J Infect*. 2020;81(2): 318-356. https://doi.org/10.1016/j.jinf.2020.04.016
6. Hourdel V, Kwasiborski A, Balière C, et al. Rapid genomic characterization of SARS-CoV-2 by direct amplicon-based sequencing through comparison of MinION and Illumina iSeq. 100TM System. *Front Microbiol*. 2020;11:11. https://doi.org/10.3389/fmicb.2020.571328
7. 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(14):3059-3066. https://doi.org/10.1093/nar/gkf436
8. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32(1):268-274. https://doi.org/10.1093/molbev/msu300
9. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14(6):587-589. https://doi.org/10.1038/nmeth.4285
10. Alm E, Broberg EK, Connor T, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill*. 2020;25(32), https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410
11. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol*. 2016;2(1):007. https://doi.org/10.1093/ve/vew007
12. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. *Virus Evol*. 2018;4(1):vex042. https://doi.org/10.1093/ve/vex042
13. Kingman JFC. The coalescent. *Stoch Process Appl*. 1982;13(3):235-248. https://doi.org/10.1016/0304-4149(82)90011-4
14. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. *Science*. 2020;369(6501):297-301. https://doi.org/10.1126/science.abc1917
15. Rambaut A, Holmes EC, O'toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020;5(11):1403-1407. https://doi.org/10.1038/s41564-020-0770-5
16. Candido DS, Claro IM, de Jesus JG, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*. 2020;369(6508):1255-1260. https://doi.org/10.1126/science.abd2161
17. Laiton-Donato K, Villabona-Arenas CJ, Usme-Ciro JA, et al. Genomic epidemiology of severe acute respiratory syndrome coronavirus 2, Colombia. *Emerg Infect Dis*. 2020;26(12):2854-2862. https://doi.org/10.3201/eid2612.202969
18. Padilla-Rojas C, Vega-Chozo K, Galarza-Perez M, et al. Genomic analysis reveals local transmission of SARS-CoV-2 in early pandemic phase in Peru. *bioRxiv*. 2020. https://doi.org/10.1101/2020.09.05.284604
19. Spada E, Sagliocca L, Sourdis J, et al. Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J Clin Microbiol*. 2004;42(9):4230-4236. https://doi.org/10.1128/JCM.42.9.4230-4236.2004
20. Ministerio de Salud, Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. Prevención y Control de Enfermedades Alerta Epidemiológica Ante La Presencia de Casos Confirmados de COVID-19 En El Perú. 2020. https://www.dge.gob.pe/portal/docs/alertas/2020/AE011.pdf
21. Facebook. Facebook Data for Good to Response to the COVID19 Pandemic. Published 2020. https://dataforgood.fb.com/docs/covid19/
22. Alteri C, Cento V, Piralla A, et al. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat Commun*. 2021;12(1):434. https://doi.org/10.1038/s41467-020-20688-x
23. Castillo AE, Parra B, Tapia P, et al. Geographical distribution of genetic variants and lineages of SARS-CoV-2 in Chile. *Front Public Health*. 2020; 8:8. https://doi.org/10.3389/fpubh.2020.562615
24. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182(4):812-827. https://doi.org/10.1016/j.cell.2020.06.043
25. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*. 2020;26:1-6. https://doi.org/10.1038/s41586-020-2895-3
26. Ramírez JD, Florez C, Muñoz M, et al. The arrival and spread of SARS-CoV-2 in Colombia. *J Med Virol*. 2021;93(2):1158-1163. https://doi.org/10.1002/jmv.26393
27. Hosseini Rad SMA, McLellan AD. Implications of SARS-CoV-2 mutations for genomic RNA structure and host microRNA targeting. *Int J Mol Sci*. 2020;21(13), https://doi.org/10.3390/ijms21134807
28. Grubaugh ND, Petrone ME, Holmes EC. We shouldn't worry when a virus mutates during disease outbreaks. *Nat Microbiol*. 2020;5(4):529-530. https://doi.org/10.1038/s41564-020-0690-4
29. Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int J Clin Pract*. 2020; 74(8):e13525. https://doi.org/10.1111/ijcp.13525
30. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *J Hum Genet*. 2020;65(12):1075-1082. https://doi.org/10.1038/s10038-020-0808-9

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Juscamayta-López E, Carhuaricra D, Tarazona D, et al. Phylogenomics reveals multiple introductions and early spread of SARS-CoV-2 into Peru. *J Med Virol*. 2021;93: 5961-5968. https://doi.org/10.1002/jmv.27167