RESEARCH ARTICLE

REVISED Genomic evidence of multiple SARS-CoV-2 introductions into Morocco [version 2; peer review: 1 approved with reservations, 1 not approved]

Bouchra Chaouni¹, Imane Allali¹, Sofia Sehli^{2,3}, Wajih Rhalem⁴,
Abdellah Idrissi Azami ^{2,3}, Nihal Habib^{2,3}, Salsabil Hamdi⁵, Lahcen Wakrim⁵,
Abderrahmane Bakkali ⁶, Mustapha Mahmoud⁷, Laila Bouguenouch⁷,
Najib Al Idrissi⁸, Youssef Bakri¹, Saaïd Amzazi¹, Said Barrijal⁹, Hassan Ghazal ²

¹Faculty of Sciences, University Mohammed V, Rabat, Morocco

²National Center for Scientific and Technical Research (CNRS), Rabat, Morocco

³Faculty of Medicine & Pharmacy, University Mohammed V, Rabat, Morocco

⁴National High School for Arts and Professions, University Mohammed V, Rabat, Morocco

⁵Institut Pasteur Morocco, Casablanca, Morocco

⁶Faculty of Medicine & Pharmacy, University Ibn Zohr, Agadir, Morocco

⁷Faculty of Medicine & Pharmacy, University Sidi Mohammed Ben Abdellah, Fes, Morocco

⁸Faculty of Medicine, University Mohammed VI for Health Sciences, Casablanca, Morocco

⁹Faculty of Sciences and Technology, University Abdelmalek Essaâdi, Tangier, Morocco

V2 First published: 06 Jul 2020, 9:679
<https://doi.org/10.12688/f1000research.24810.1>
Latest published: 24 Aug 2020, 9:679
<https://doi.org/10.12688/f1000research.24810.2>

Abstract

Background: The recent emergence of a novel coronavirus (SARS-CoV-2) has caused serious public health concerns due to its rapid dissemination worldwide. A total of 8,931 positive cases had been reported in Morocco by the 16th of June 2020.

Methods: To better understand the SARS-CoV-2 pandemic in this North African country, we analyzed the complete genome sequences of the virus related to Morocco by constructing a phylogenetic tree and creating a variant network using the available Moroccan and other sequences in dedicated databases.

Results: Phylogenetic and variant network analyses of SARS-CoV-2 strains from early patients with COVID-19 in Morocco showed multiple spatiotemporal introductions from Italy (ten), France (seven), Spain (one) and Portugal (one). This is consistent with the assumption that the early infections in Morocco were imported, mainly from Europe. The 17 virus strains form two independent phylogenetic clusters and provide evidence for early community-based transmission following the initial introductions of the virus. We then catalogued 13 novel mutations in the SARS-CoV-2 isolates from Moroccan patients. Interestingly, the recurrent missense variant A>G at position 23,403 in the spike gene, known to be associated with virus severity, has been

Open Peer Review

Approval Status ? X

	1	2
version 2 (revision) 24 Aug 2020		 view
version 1 06 Jul 2020	 view	

1. **Keith Crandall** , The George Washington University, Washington, USA

2. **Richard L Tillet** , University of Nevada Las Vegas, Las Vegas, USA

Any reports and responses or comments on the article can be found at the end of the article.

identified in all Moroccan isolates.

Conclusions: These primary findings testify of the importance of the genomic surveillance strategies as a means of understanding the virus spread dynamics, counteracting the pandemic outbreak, and drawing useful lessons for dealing with any future emerging infectious pathogens.

Keywords

Coronavirus, SARS-CoV-2, Covid-19, Morocco, Molecular Epidemiology, Genomics Surveillance, Phylogenetics, Variant Network Analysis

This article is included in the **Emerging Diseases and Outbreaks** gateway.

This article is included in the **Coronavirus** collection.

Corresponding author: Hassan Ghazal (hassan.ghazal@fulbrightmail.org)

Author roles: **Chaouni B:** Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Allali I:** Conceptualization, Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing; **Sehli S:** Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation; **Rhalem W:** Formal Analysis, Investigation, Methodology, Software, Visualization; **Idrissi Azami A:** Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization; **Habib N:** Data Curation, Formal Analysis, Investigation, Methodology, Visualization; **Hamdi S:** Validation, Writing – Review & Editing; **Wakrim L:** Conceptualization, Writing – Review & Editing; **Bakkali A:** Conceptualization, Methodology, Writing – Review & Editing; **Mahmoud M:** Writing – Review & Editing; **Bouguenouch L:** Conceptualization, Methodology, Writing – Review & Editing; **Al Idrissi N:** Conceptualization, Methodology, Project Administration, Writing – Review & Editing; **Bakri Y:** Conceptualization, Supervision, Writing – Review & Editing; **Amzazi S:** Conceptualization, Writing – Review & Editing; **Barrijal S:** Conceptualization, Methodology, Writing – Review & Editing; **Ghazal H:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2020 Chaouni B *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Chaouni B, Allali I, Sehli S *et al.* **Genomic evidence of multiple SARS-CoV-2 introductions into Morocco [version 2; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2020, 9:679 <https://doi.org/10.12688/f1000research.24810.2>

First published: 06 Jul 2020, 9:679 <https://doi.org/10.12688/f1000research.24810.1>

REVISED Amendments from Version 1

This article has been updated to remove references to data that was in breach of GISAID's terms and conditions. The data availability statement details how users can access the data directly from GISAID. The article now includes acknowledgment of the contributions of both the Submitting and the Originating laboratories of the GISAID data used in this study.

Any further responses from the reviewers can be found at the end of the article

Introduction

Following its onset in December 2019, several cases of a new respiratory illness were reported in the city of Wuhan, Hubei province, China¹. The disease, named coronavirus disease 2019 (COVID-19), was later on confirmed as caused by a novel coronavirus that was subsequently called the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2)². On the 12th of March 2020, the ongoing SARS-CoV-2 outbreak was declared as a pandemic by the World Health Organization (WHO)³. As of June 16th 2020, there have been 8,228,025 confirmed cases with 444,442 deaths around the world (John Hopkins Center (JHC))⁴. Currently, Morocco has reported 8931 confirmed cases and 212 deaths associated with COVID-19 (JHC update of June 16th 2020). To gain further understanding of the molecular epidemiology of the outbreak in Morocco, we conducted a phylogenetic and a Variant Network Analyses of the full-genome sequences of 21 SARS-CoV-2 strains, 19 were isolated from Covid-19 patients in Morocco, one sequence was isolated from a patient in Melilla and another one was isolated from a Moroccan patient in Cadiz, Andalusia.

Methods

Data set

The Morocco related-sequences used in this study have been deposited in the GISAID database⁵ by the Laboratory of the Royal Gendarmerie of Morocco (LRAM) (one sequence), the Pasteur Institute of Morocco (IPM) (17 sequences), the Anoual Laboratory in Casablanca (one sequence), and the SeqCovid Spanish Project (two sequences). The sequences were collected from the GISAID database and analyzed in comparison with 500 selected genomes, that were collected between the 15th of February 2020 (2 weeks before the first detected case in Morocco) and the 30th of March 2020 (2 weeks after the international borders of Morocco were closed; lockdown). Only high-quality sequences have been included in this study.

Variant calling and network analysis

The Moroccan sequences were mapped to the Wuhan reference genome (NC_045512) using Bowtie 2.3.5.1⁶. Variants were called using mpileup 1.7 and bcftools 1.9⁷. The variants were annotated using the China National Center for Bioinformatics Annotator. Variant network analysis was performed using Gephi 0.9.2.

Phylogenetic analysis

A total of 500 sequences from different countries were retrieved from the GISAID database (www.gisaid.org)⁵ and aligned

using Muscle 3.8.1551 multiple sequence alignment⁸. A maximum likelihood model was created using RaxML 8.2.12⁹ with a bootstrap of 100 using the Wuhan reference genome (NC_045512) as outgroup. The phylogenetic tree was generated using Figtree 1.4.4¹⁰.

Results and discussion

Phylogenetic tree analysis

To put the complete genomes of SARS-CoV-2 that were isolated from 19 Moroccan patients, a patient from Melilla and a Moroccan patient from Cadiz (Andalusia, Spain), into the context of the global pandemic, they were aligned together with the dataset of 500 SARS-CoV-2 complete genomes from different countries. Extended data, Figure 1¹¹ shows the estimated maximum likelihood phylogeny. To have a good presentation of the phylogeny, we created a circular presentation subset of 120 sequences from our dataset (collected from 18th of February 2020 (15 days before the first Moroccan case) to 15th of March 2020 (Date of closing international borders in Morocco)) (Figure 1b).

The extended phylogenetic tree is subdivided into seven clades that correspond to the main seven SARS-CoV-2 strain types GR (red), GH (green), G (yellow), S (pink), V (magenta), L (light gray) and O (others; dark gray). The four strain types S, G, GR, and GH are scattered all over the world (Extended data, Figure 2)¹¹. According to GISAID (update of the 12th of June 2020), they made their first appearance in February 2020, being G the strain type that gave origin to the GH and GR ones, while the S strain was mainly found in Spain¹².

Zooming into the trees (Extended data, Figure 1)¹¹, Figure 1a shows how the Moroccan isolates group into four independent clusters. In the GR clade, the sequences GISAID IDs: EPI_ISL_459973, EPI_ISL_459966 and EPI_ISL_459965 (the first infected patient in Morocco) cluster together with sequences from Italy (GISAID ID: EPI_ISL_417922), Thailand (GISAID ID: EPI_ISL_430837), Mexico (GISAID ID: EPI_ISL_452139), and Iceland (GISAID ID: EPI_ISL_417829) (Figure 1a: A1–3). Interestingly, the hosts of these four sequences were all travelling in Italy during the same period. This suggests that the first Moroccan patient (GISAID: EPI_ISP_459965), that returned to Morocco from recent travel to Bergamo, Lombardy (the most affected region in Italy), and the three other individuals from Mexico, Iceland and Thailand (GISAID: EPI_ISL_452139, EPI_ISL_417829 and EPI_ISL_430837, respectively) were probably infected by the same person in Italy (GISAID: EPI_ISL_417922). From this cluster, we infer, based on the sample collection dates, that the first Moroccan patient (GISAID: EPI_ISP_459965) probably infected the two other Moroccan patients whose infecting virus sequences are found in the same cluster (GISAID: EPI_ISL_459973 and EPI_ISL_459966). The Moroccan sequences (GISAID IDs: EPI_ISL_458150, EPI_ISL_459980 and EPI_ISL_459984) that cluster together in the GR clade, are close to sequences from Iceland (GISAID IDs: EPI_ISL_417861, EPI_ISL_417863 and EPI_ISL_417852) whose carriers have travel history to Italy. Furthermore, all these three Icelandic sequences are related to one Italian sequence (GISAID ID: EPI_ISL_460081). For their

Figure 1. Phylogenetic tree analysis. (a) Maximum likelihood phylogeny, constructed using complete SARS-CoV-2 genomes from the GISAID database⁵. The blue sequences represent the Moroccan isolates (19), a strain isolated from Melilla and a strain isolated from a Moroccan patient in Cadiz, Andalusia. The phylogenetic tree is divided into four clades: the red part for the GR strains, the green one for the GH strains, the yellow one for the G strains, and the pink one for the S strain. This Figure represents selected sub-tree extracts from the extended phylogenetic tree in *Extended data, Figure 1*¹¹. A1–3: represent clusters of Moroccan strains originated from Italy and belonging to the GR clade; B: represents a cluster of Moroccan strains from a French origin and belonging to the GH clade; C: represents the Moroccan strain isolated from Ouarzazate, it originated from Portugal and belongs to the G clade; D: represents a cluster of Moroccan strains originated from France and belonging to the G clade; E: represents a Moroccan strain that originated from Italy and it belongs to the G clade; F: represents a Moroccan strain from a Spanish origin and belonging to the G clade, G: represents the sequence isolated from Melilla belonging to the G clade with a Spanish origin, H: represents the sequence isolated from Cadiz; it belongs to the S clade and it's close to strains from Spain. (b) Circular representation of the phylogenetic tree of a subset of 120 sequences collected before closing Moroccan international borders, constructed using complete SARS-CoV-2 genomes from the GISAID database⁵. The blue sequences represent the Moroccan isolates (19). The phylogenetic tree is divided into three clades: the red part for the GR strains, the green one for the GH strains, and the yellow one for the G strains. The clade S containing the Cadiz sequence is absent as we removed this sequence from this phylogenetic analysis.

part, the Moroccan sequences EPI_ISL_459983, EPI_ISL_459977 and EPI_ISL_459978 cluster with another sequence from Italy (GISAID ID: EPI_ISL_452186). Surprisingly, many sequences from the GR clade were isolated from patients with known travel history to Italy (Figure 1a: A1–3). Therefore, we can conclude that all the viral sequences in this clade, where 9 Moroccan viral sequences clustered, were of Italian origin.

The four Moroccan sequences that cluster together within the G type strain (GISAID IDs: EPI_ISL_459968, EPI_ISL_459974, EPI_ISL_459972 and EPI_ISL_459981) were close to French sequences (GISAID ID: EPI_ISL_417333, EPI_ISL_443270 and EPI_ISL_416752). These four Moroccan sequences were collected the 17th of March, the 20th of March, the 20th of March and the 19th of April 2020, respectively. Interestingly, the third Covid-19 case in Morocco was a French male tourist, in his 30s that arrived in Morocco on the 7th of March 2020 as

reported by the Moroccan Health authorities. He, therefore, was most likely the source of the Moroccan viral strains that belong to the G type. The case reported on the 19th of April 2020 (sequence with the GISAID ID: EPI_ISL_416752) seems to be the result of a local community transmission.

The Moroccan strain isolated from Ouarzazate (South of Morocco) (GISAID ID: EPI_ISL_451400) is close to the Portuguese sequences (GISAID ID: EPI_ISL_413648 and EPI_ISL_453947) in the G type suggesting a third introduction path to Morocco from Portugal. For their part, the Moroccan sequence (GISAID ID: EPI_ISL_459967), that also belongs to the G type, is close to the Spanish sequences (GISAID IDs: EPI_ISL_455349 and EPI_ISL_455332) making Spain another point of introduction to Morocco. Furthermore, the Moroccan sequence (GISAID ID: EPI_ISL_459975) is closely related to the Italian G strain (GISAID: EPI_ISL_417921),

which represents another Italian introduction to Morocco. It's also similar to a sequence from Iceland (GISAID ID: EPI_ISL_417730), where the host patient has travel history to Italy (Figure 1a: E) and could have been either infected in Italy or by another asymptomatic carrier coming from there.

Three other viral sequences isolated from Moroccan patients (GISAID IDs: EPI_ISL_459976, EPI_ISL_459982 and EPI_ISL_459979) appear in the clade of the GH strain types where they cluster with French sequences (GISAID IDs: EPI_ISL_416748 and EPI_ISL_416501) (Figure 1a: B) suggesting a further introduction to Morocco, once more from France.

The viral sequence isolated from a patient in Melilla (GISAID ID: EPI_ISL_455344) is clustered within the G type, with Spanish sequences (GISAID IDs: EPI_ISL_419235, EPI_ISL_450337 and EPI_ISL_419236). While the sequence isolated from a Moroccan patient in Cadiz, Andalusia (GISAID ID: EPI_ISL_452463) (Figure 1a: G) that belongs to the S type, also clusters with Spanish sequences; however, this sequence is quite far from the other Spanish related sequences. Thus, supporting the apparent neutrality of the host's genetics for the transmission and potential selection on the type of strain.

Finally, similarity analysis of the Moroccan sequence (GISAID ID: EPI_ISL_467299) collected on May 21st, 2020 and sequenced by the LRAM Laboratory, is close to the first identified French cluster, suggesting that the virus strain circulating in Morocco did not experience any major mutations and, once again, confirms the local community-based transmission of the virus. This sequence was not included in the overall analysis, as it was deposited when the writing of this work was finalizing.

Thus, the phylogenetic analysis shows that the virus outbreak in Morocco was likely the result of multiple introductions. We can highlight four independent introductions to Morocco from Italy, France, Spain, and Portugal. This finding has obvious implications for the epidemiological tracing of the pathogenic agent's initial introduction that caused the current outbreak in Morocco. Generally, the viral sequences isolated from Moroccan patients showed close relationships primarily to European strains. The geographic nearness and tourist and migratory connections, therefore, play a key role in the spread of the virus. We caution that further analyses are needed to evaluate the statistical robustness of the interference suggested herein.

Network analysis

To further understand the evolution of the SARS-CoV-2 virus within the Moroccan population and trace the infection pathways, we performed a Variant Network Analysis using Gephi. We used the complete sequences of the 21 SARS-CoV-2 genomes, 19 from Morocco, and two sequences from Melilla and Andalusia (Figure 2). The variant network in Figure 2 shows three main clusters, one isolated sequence from Portugal and another one from Cadiz sharing no variants with the other

sequences. The first cluster comprising the following viral strains, all closely related to each other and diagnosed in Casablanca (west Morocco) (GISAID ID: EPI_ISL_458150), IPM_1 (GISAID ID: EPI_ISL_459965), IPM_2 (GISAID ID: EPI_ISL_459966), IPM_13 (GISAID ID: EPI_ISL_459980), IPM_15 (GISAID ID: EPI_ISL_459978), IPM_7 (GISAID ID: EPI_ISL_459973), IPM_16 (GISAID ID: EPI_ISL_459977), IPM_10 (GISAID ID: EPI_ISL_459984) and IPM_17 (GISAID ID: EPI_ISL_459983) were closely related to each other. This is in agreement with the phylogenetic tree where all these sequences belong to the GR clade and appear to have been originated from Italy (Figure 1 and Extended data, Figure 1)¹¹. The second cluster (Blue) made out of the sequences labeled IPM_5 (GISAID ID: EPI_ISL_459974), IPM_6 (GISAID ID: EPI_ISL_459968), IPM_8 (GISAID ID: EPI_ISL_459972) and IPM_14 (GISAID ID: EPI_ISL_459981) belong to the clade G and has a French sequence (GISAID ID: EPI_ISL_418222) as a root in the phylogenetic tree. The sequences IPM_4 (GISAID ID: EPI_ISL_459976), IPM_11 (GISAID ID: EPI_ISL_459982) and IPM_12 (GISAID ID: EPI_ISL_459979) belong to the GH clade and seem to have originated from another French sequence (GISAID: EPI_ISL_418219). Interestingly, this sequence belongs to the G clade and is close to a strain that was isolated from an Icelandic patient that was most probably infected during his stay in Italy (GISAID ID: EPI_ISL_417730). Another separated Moroccan sequence, labelled IPM_3 (GISAID ID: EPI_ISL_459967), is closer to Spanish sequences (GISAID ID: EPI_ISL_455349, EPI_ISL_455332). The sequence of the viral strain from Ouarzazate (GISAID ID: EPI_ISL_451400) is closely related to a Portuguese sequence (GISAID ID: EPI_ISL_413648), just as shown in the phylogenetic tree. Yet, all these sequences belong to the G clade, which has an Italian sequence as a root. Hence, the Variant Network Analysis is in accordance with the phylogenetic tree and further supports our deduction of multiple SARS-CoV-2 introductions to Morocco from at least four European countries. Thus, while geography marks the variation that the virus undergoes in a way that we can identify geographically separated strains, population interconnection, via travel and migratory flows, seems more important than the geographical nearness for the worldwide spread of the virus.

To further test the multiple introductions to Morocco from the aforementioned four European countries, we performed a Variant Network analysis using all the Moroccan sequences as one group and compared them with randomly selected sequences from different countries (Italy, France, Spain and Portugal, in addition to Austria, Australia, and Brazil). These results show that the sequences detected during the early stages of the epidemic in Moroccan share variants with Italian (seven variants), French (four), Portuguese (three), and Spanish (three) strains. They also share one variant with strains from Austria, one with Brazil and another one with Australia (Figure 3). Interestingly, there was one shared variant between all countries located in the Spike gene (23403A>G) which induces the amino acid change D614G; it began spreading in Europe in early February and, when introduced to new regions, it rapidly became the dominant form. This mutation is suspected to increase the transmissibility of the virus¹³.

Figure 2. Variant network analysis of the 19 Moroccan genome sequences, one sequence from Melilla and another one from a Moroccan host in Cadiz, Andalusia. Every dot represents a sequence, and each edge on the network connects, using a shared variant, a sequence to one or more other sequences. The red circle group Moroccan sequences originated from Italy, the green one group sequences originated from France, the purple one assemble sequences originated from Spain, and the blue one represents the sequence originated from Portugal; we used different color gradients to separate sequences that belong to different clades in the same cluster. A: Strain isolated from Cadiz, it belongs to the S clade and it shares no variants with Moroccan strains, including the Spanish ones. M: represents the strain isolated from Melilla, it belongs to the G clade in a cluster of Spanish sequences. IPM3: is a Moroccan strain isolated by Pasteur Institut of Morocco (IPM), it belongs to the G clade and it's coming from Spain. O: represents the strain isolated by LRAM from Ouarzazate, it's the only sequence with a Portuguese origin and it belongs to the G clade. IPM9: represents a sequence collected by IPM, it originated from Italy and belongs to G clade. C; IPM1; IPM2; IPM7; IPM10; IPM13; IPM15-17: represent a strain isolated by Anoual laboratory in Casablanca (C) and sequences collected by IPM, they belong to the GR clade and they all are from Italian origin. IPM4; IPM11-12: represent strains isolated by IPM, they originated from France and they belong to the GH clade. IPM5-6; IPM8; IPM14: represent strains isolated by IPM they are from a French origin and they belong to the G clade.

Variant calling

The genome-wide SNPs and the corresponding amino-acid positions and variations of the virus proteins are described in Table 1. Our results showed that all 20 sequences including the sequence coming from the Melilla patient, shared four mutations in common and 13 novel mutations exclusively present

in the 19 sequences isolated in Morocco (Table 2). The four common mutations happened to be the same recurrent mutations described in several reports¹⁴. The mutation in the leader sequence (241C>T) is one of the most common mutations in the SARS-CoV-2 genome; it affects an important genomic site for discontinuous sub-genomic replication. The mutation in the

Figure 3. Variant network analysis elucidating the shared variants between sequences from different countries. Each dot represents a group of 19 sequences randomly selected from each country. The Moroccan node represents all 19 sequences isolated in Morocco. Morocco shares more variants with Italy (seven), France (four), Portugal (three) and Spain (three) than with Brazil (one), Austria (one) and Australia (one).

Table 1. List of the common mutations in all 19 sequences of SARS-CoV-2 from Morocco that are shared with European strains.

Mutation	Position	Genomic region	Type of mutation
C>T	241	5'UTR	upstream_gene_variant
C>T	3037	ORF1ab	Synonymous_variant
C>T	14408	ORF1ab	Missense_variant
A>G	23403	Gene S	Missense_variant

5'UTR genomic region co-evolved with two other mutations, 14408C>T and 23403A>G. They both affect critical RNA replication proteins (241C>T, 14408C>T) and the ACE2 receptor binding protein S (23403A>G). We noticed that these four mutations are prevalent in the virus isolates from Europe, where the infections seem to be more severe. In fact, Yin *et al.*¹⁵ suggested that these mutations could increase the transmissibility of the virus. The sharing of these mutations between European and Moroccan isolates further indicates that strains circulating in Morocco came mainly from Europe. Moreover, at least one variant of the Moroccan sequences of SARS-CoV-2 is shared with one of the other countries (Table 3). The

Table 2. List of the mutations exclusively found in the 19 Moroccan sequences of SARS-CoV-2.

Mutation	Position	Genomic region	Type of mutation
C>T	3874	ORF1ab	Synonymous variant
C>T	5497	ORF1ab	Synonymous_variant
C>T	8175	ORF1ab	Missense_variant
C>T	12100	ORF1ab	Synonymous_variant
G>T	15380	ORF1ab	Synonymous_variant
C>T	18508	ORF1ab	Missense_variant
C>T	21513	ORF1ab	Synonymous_variant
C>T	26559	Gene M	Missense_variant
A>C	691	ORF1ab	Synonymous_variant
G>A	198	5'UTR	Upstream gene variant
T>A	28738	Gene M	Synonymous_variant
T>A	4011	ORF1ab	Missense_variant
C>T	313	ORF1ab	Synonymous_variant

Table 3. List of Moroccan (M: Morocco) variants shared with at least one of the other studied countries (F: France; S: Spain; P: Portugal; I: Italy; A: Australia; Au: Austria; B: Brazil).

Mutation	Position	Countries	Type of mutation	Genomic region
C>T	241	M, F, I, P, S	Upstream_gene_variant	5'UTR
C>T	3037	M, F, I	Synonymous_variant	ORF1ab
C>T	14408	M, F, I, P, S	Missense_variant	ORF1ab
A>G	23403	M, F, I, P, S, B, A, Au	Missense_variant	Gene S
G>C	28883	M, I	Missense_variant	Gene-N
G>A	28881	M, I	Missense_variant	Gene-N
G>A	28882	M, I	Synonymous_variant	Gene-N

mutations found in the sequences identified in Morocco, and the identification of amino acid change, should be further investigated in order to understand whether they affect the transmission and clinical characteristics of the SARS-CoV-2 virus circulating in Morocco.

Conclusions

Phylogenetic and variant network analyses of SARS-CoV-2 sequences from early patients with COVID-19 in Morocco showed multiple spatiotemporal introductions, primarily from

Italy (ten), France (seven), Spain (one) and Portugal (one). The results also provide evidence for early community-based transmission. A variant calling analysis allowed us to catalogue new mutations in SARS-CoV-2 isolates from Morocco. Interestingly, the recurrent missense variant A>G at position 23,403 in the spike gene known to be associated with virus severity has been identified in all Moroccan isolates. However, the lack of demographic and clinical information on most of the sequences of the Moroccan isolates deposited in the database prevented us from inferring the potential link between the mutations and

clinical effects of the strains. Also, the number of Moroccan genomes available at the time of this analysis hampers robust statistical inferences. Still, the results of the present work offer interesting insights into how the virus got into Morocco and primary lessons as to understanding the dynamics of the initial introductions and local transmission of SARS-CoV-2 in Morocco. This being a global pandemic, the conclusions of this study, taken together with those of recent similar studies such as those done in Egypt¹⁶, South Africa¹⁷ and Brazil¹², might be applicable to other developing countries. The results of the methodology of genomic surveillance that we developed, and the methodology itself would be helpful in the case of resurgence or the emergence of any new pandemic-causing pathogen. Integrative analysis of SARS-CoV-2 should promote our understanding of the virus dynamics and interactions with hosts and environments. The method can even be further extended to reconstruct the evolutionary origins of the enhanced pathogenicity of SARS-CoV-2 and other coronaviruses that are severe human pathogens¹⁸.

Data availability

Source data

The datasets are available from GISAID⁵ (<https://www.gisaid.org/>; GISAID accession numbers are presented in the Acknowledgment Table, see Extended data¹¹). Access to the data requires registration and agreement to the conditions for use at: <https://www.gisaid.org/registration/register/>.

Extended data

Zenodo: Genomic evidence of multiple SARS-CoV-2 introductions into Morocco. <http://www.doi.org/10.5281/zenodo.3964418>¹¹.

This project contains the following extended data:

- **Extended data Figure 1** (PDF). (Full size image of the phylogenetic tree seen in Figure 1.)
- **Extended data Figure 2** (PNG). (Full genome tree derived from all outbreak sequences.)
- **Acknowledgment Table** (PDF). (Acknowledgment Table of the sequences assessed in this study.)

All supplementary material and figures can be found at GitHub: https://github.com/covidmor/covid19_morocco.

Extended data are available under the terms of the **Creative Commons Zero “No rights reserved” data waiver** (CC0 1.0 Public domain dedication).

Acknowledgements

We acknowledge the contributions of both the Submitting and the Originating laboratories of the GISAID data⁵ used in this study, as listed in the Acknowledgment Table (see Extended data¹¹).

References

1. Wu P, Hao X, Lau EHY, *et al.*: **Real-time tentative assessment of the epidemiological characteristics of novel coronavirus infections in Wuhan, China, as at 22 January 2020.** *Euro Surveill.* 2020; **25**(3): 2000044. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Uddin M, Mustafa F, Rizvi TA, *et al.*: **SARS-CoV-2/COVID-19: Viral Genomics, Epidemiology, Vaccines, and Therapeutic Interventions.** *Viruses.* 2020; **12**(5): 526. [cited 2020 Jun 15]. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. **WHO announces COVID-19 outbreak a pandemic.** 2020; [cited 2020 Jun 15]. [Reference Source](#)
4. Johns Hopkins Center: **Coronavirus COVID-19 (2019-nCoV).** 2020; [cited 2020 Jun 16]. [Reference Source](#)
5. Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** *Glob Chall.* 2017; **1**(1): 33–46. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res.* 2004; **32**(5): 1792–7. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Stamatakis A: **RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics.* 2006; **22**(21): 2688–90. [PubMed Abstract](#) | [Publisher Full Text](#)
10. Rambaut A: **FigTree Version 1.4.4.** 2020. [cited 2020 Jun 15]. [Reference Source](#)
11. covidmor: **covidmor/covid19_morocco: Genomic evidence of multiple SARS-CoV-2 introductions into Morocco (Version 1.0).** *Zenodo.* 2020. <http://www.doi.org/10.5281/zenodo.3964418>
12. de Jesus JG, Sacchi C, Claro I, *et al.*: **First cases of coronavirus disease (COVID-19) in Brazil, South America.** 2020. [Reference Source](#)
13. Korber B, Fischer WM, Gnanakaran S, *et al.*: **Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2.** *bioRxiv.* 2020; 2020.04.29.069054. [Publisher Full Text](#)
14. Pachetti M, Marini B, Benedetti F, *et al.*: **Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant.** *J Transl Med.* 2020; **18**(1): 179. [cited 2020 Jun 4]. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Yin C: **Genotyping coronavirus SARS-CoV-2: methods and implications.** *Genomics.* 2020; **112**(5): 3588–3596. [cited 2020 Jun 15]. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Kandeil A, Mostafa A, El-Shesheny R, *et al.*: **Coding-Complete Genome Sequences of Two SARS-CoV-2 Isolates from Egypt.** *Microbiol Resour Announc.* 2020; **9**(22): e00489–20. [cited 2020 Jun 15]. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Mavian C, Pond SK, Marini S, *et al.*: **Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-CoV-2 infections unreliable.** *Proc Natl Acad Sci U S A.* 2020; **117**(23): 12522–12523. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Gussow AB, Auslander N, Faure G, *et al.*: **Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses.** *bioRxiv.* 2020; 2020.04.05.026450. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ✕

Version 2

Reviewer Report 02 February 2021

<https://doi.org/10.5256/f1000research.28952.r77070>

© 2021 Tillett R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Richard L Tillett

Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, Las Vegas, Nevada, USA

Summary: The authors describe the genomes of multiple SARS-CoV-2 strains isolated from or related to Morocco. They compare them with extant SARS-CoV-2 genomes from around the globe and present evidence for multiple introductions, using phylogenies, network analyses and variant analyses.

Caution required on phylogenetic conclusions:

Overall, I think the work needs to be more cautious of overinterpretation of the phylogenetic neighborhoods and network placements, as this risks drawing stronger conclusions than may be strictly warranted. I also wish to see additional descriptions of the sequence and variant evidence that distinguish the specific introductions identified. In present form, the variant analysis in the manuscript clearly supports at least two introductions to Morocco: at least one G-clade introduction, and at least one GR-clade introduction. Variant analyses demonstrating GH introductions should be similarly included, as well as an elaboration additional variants which may support multiple introductions of particular clades. There may well be cases with identical genotypes, but contrasting travel histories that would support multiple introductions, and I would rely on these sets of facts (genotypes and histories) more heavily than on the networks, in drawing these conclusions. I would like to see changes in the manuscript to support claims of introductions with particular variants, where possible, and to verify that sequences of given clades have the expected variants for said memberships and relationships to extant sequences.

I caution the authors against making exact claims of exact transmission events based on 500 sequences without strong evidence from contact tracing. With some notable exceptions (recently identified B.1.1.7 variant, for example), the mutation rate in SARS-CoV-2 is thought to be low, at just 1 mutation per two weeks, in a given transmission chain. Thus, many persons in transmission chains or clusters, including those whose genomes were never sequenced, will harbor highly similar viral genotypes. Unless they are able to reconstruct the transmission via contact tracing, I do not think either *tree models* or *sequence similarity* are sufficient to make declarations such as:

This suggests that the first Moroccan patient (GISAID: EPI_ISP_459965), that returned to Morocco from recent travel to Bergamo, Lombardy (the most affected region in Italy), and the three other individuals from Mexico, Iceland and Thailand (GISAID: EPI_ISL_452139, EPI_ISL_417829 and EPI_ISL_430837, respectively) were probably infected by the same person in Italy and we infer, based on the sample collection dates, that the first Moroccan patient (GISAID: EPI_ISP_459965) probably infected the two other Moroccan patients whose infecting virus sequences are found in the same cluster (GISAID: EPI_ISL_459973 and EPI_ISL_459966)

Statements like those I have quoted are risky and may over-interpret the data. How do we know that these four persons were infected by the same person in Italy? How do we know that one patient in Morocco infected the two others?

Another example I find risky in its conclusion is this:

Therefore, we can conclude that all the viral sequences in this clade, where 9 Moroccan viral sequences clustered, were of Italian origin.

I do not agree with this conclusion. Clade GR was not and is not unique to Italy. Multiple introductions could have occurred quite easily. The sequences with travel histories linking them to Italy have good evidence that they acquired it from Italy. The others could have obtained GR by the known introductions or by unknown introductions of GR-type SARS-CoV-2. The presence or absence of additional variants could be the clue for such inferences. (Or, given the low mutation rate of SARS-CoV-2, there might not be additional mutations from which to infer).

Similarly, the description of Figure 2 makes the declarative statement that *"IPM3: is a Moroccan strain isolated by Pasteur Institut of Morocco (IPM), it belongs to the G clade and it's coming from Spain."* This should be replaced with a more neutral description of the evidence, rather than stating the conclusion.

Variants support needed for conclusions:

I find the mutations listed in Tables 3 hard to reconcile with the findings of transmission and networks presented elsewhere. The variants shown in Table 3 identify clades G and clades GR only, with no additional mutations in support of either GH or multiple, specific introductions of G and/or GR. Previous to the variant analysis section, the authors present networks that identify similarities between individual sequences and sets of Moroccan sequences with individual and sets of sequences from GISAID, and use these similarities to draw lines of inferences of introductions. My presumption would be that within these networks there would be variants held in common amongst neighbors. Indeed, in Table 3 they list the quartet of variants common to clade G and descendants (241, 3037, 14408, and 23403), found in many (or all? I believe most. More on that in a moment) of the Moroccan sequences and in many of the 500 GISAID sequences. Table 3 also lists a triplet of mutations (28881-28883) found in some of their sequences, also found in Italian sequences from their GISAID set, and also known to me to be found in essentially all of clade GR. But the table only includes these 7 sites and thus seems incomplete to me. Again, I presume that sequences of inferred relationship would be supported by additional mutations beyond these. For example, the GH and S introductions stated previously in the text – ought not the sequence evidence for these be presented on this table too? S clade, in particular, also carries with it the presumption of a lack of the clade G mutations, and the most iron-clad evidence for S clade membership would be C8782T and/or T28144C (along with a lack of the G mutations). Similarly, GH membership is best supported by all or most of the G variants plus G25563T. See the

text in the upper-left corner of Figure 1 in Rabaan et al., 2020 for these explicit listings <https://ann-clinmicrob.biomedcentral.com/articles/10.1186/s12941-020-00384-w>. See also this statement from GISAID containing similar nomenclature notes <https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/>.

Generally speaking, I do wish to see evidence of these (or other) markers in support of clade inferences. Travel histories can augment this evidence to argue for multiple introductions within a clade as well, especially for a broad clade like G, but without presenting the evidence for GH markers (or some other markers) how can one argue for the tree neighborhood demonstrating the introduction of GH into Morocco?

If certain markers don't exist in particular Moroccan sequences, in common with the global sequences they are being related to, then it raises the risk that the phylogenies and networks are creating the appearance of relationships between isolates that lack support on the genomic level, or could possibly even be artifactual. I assume that this evidence does exist, and the content in the Variant Analysis section should be amended to include all of it to provide sequence-level support for the relationships posited by the trees and networks.

I also wish to see that level of sequence support for variants shared between the four Moroccan G sequences and the French G sequences described on page 4. What evidence is there that separates these G sequences from the G sequences from Ourzazate and Portugal? And what distinguishes each of those groups from the G sequences linking Morocco and Spain? Also, at the end of the paragraph discussing French sequences, it appears that the case of community transmission is mislabeled. The sentence reads:
The case reported on the 19th of April 2020 (sequence with the GISAID ID: EPI_ISL_416752) seems to be the result of a local community transmission.

However, that GISAID ID is one of the IDs of the French sequence listed previously in the paragraph. (Every ID in the manuscript should probably be double-checked for accuracy. I only noticed this one by happenstance).

Additional notes:

I also don't understand the authorial intent of the following statement on page 5
however, this sequence is quite far from the other Spanish related sequences. Thus, supporting the apparent neutrality of the host's genetics for the transmission and potential selection on the type of strain.

Which host genetics are they referring to? This seems very non-sequitur. We know that SARS-CoV-2 strains infect humans without too much regard for our individual genetics. The authors should remove this statement or clarify their intent.

If the authors have evidence that the D614G clades (G and descendants) result in more severe COVID, they should cite it. Otherwise, the statements (page 7)

We noticed that these four mutations are prevalent in the virus isolates from Europe, where the infections seem to be more severe.

And page 8

Interestingly, the recurrent missense variant A>G at position 23,403 in the spike gene known to be associated with virus severity has been identified in all Moroccan isolates.

should be modified to merely remark on the consensus of increased transmissibility of G-clade.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research background includes sequencing technology, bioinformatics, and SARS-CoV-2 variant analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Version 1

Reviewer Report 13 July 2020

<https://doi.org/10.5256/f1000research.27373.r66513>

© 2020 Crandall K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Keith Crandall

Computational Biology Institute, Milken Institute School of Public Health, The George Washington University, Washington, DC, 20052, USA

Summary: This paper reports on new sequence data on the SARS-CoV-2 outbreak in Morocco.

Through phylogenetic analyses they identify 4 distinct introductions to Morocco from distinct European origins followed by local transmissions. The analysis are OK, but need some additional work (see below). Specifically, there are no bootstrap analyses discussed to provide confidence in the resulting phylogenetic relationships (the main result of the paper). The authors use an excellent sampling strategy, focusing on Moroccan sequences and then those available from two weeks before the first known case in Morocco and two weeks after the closing of the borders. Their data are deposited with GISAID and thereby publicly available (with some effort and approval), but they also provide their entire working set on GitHub. Excellent! This is the way to do open science which is so critical for the COVID studies. However, the nomenclatural approach is confusing and the phylogenetic presentation could use some help. There are a lot of different things going on with this analysis (lots of different metadata to focus on) and the authors could do a better job of making things more clear in their presentation. Below are some ideas to help with both the analyses and interpretation.

Abstract/Results spike gene variant associated 'with virus severity' – this doesn't make sense to me. What does it mean to have a severe virus? Do you mean severity in clinical outcomes?

Methods/Data Set – you say 500 sequences were selected from GISAID both before and after the initial case in Morocco. I cannot, however, get a feel for the relative split or the relative timeline. It would be really helpful to have a graph that had time along the x-axis (perhaps by week or something) and then different colors for the pre-GISAID, post-GISAID, and Moroccan sequences.

Methods/Phylogenetic analysis – the authors state the 'model was created using RaxML'. First, RAXML doesn't create models. It estimates phylogenies. Second, it needs a model to do this. What model was used in this analysis? There is an updated version of ModelTest, ModelTest-NG <https://academic.oup.com/mbe/article/37/1/291/5552155> that helps with optimizing a model of evolution for a data set to then feed into RAXML. I recommend the authors use this tool to justify a model choice, report the model, and then use it for the ML estimate. Then 'FigTree' doesn't 'generate' a tree, it is a tool for phylogeny visualization. I would recommend iTOL as an alternative. iTOL also allows for more legible mapping of metadata on the phylogeny: <https://itol.embl.de/> Given that the phylogeny seems to be the main result, the authors should conduct a bootstrap analysis with at least 1000 replicates and show bootstrap support values on the phylogeny as an assessment of confidence.

In general, no details of the analyses are provided and it is therefore impossible to replicate the study. The authors simply list software. They need to be more explicit about how they actually ran their analyses.

Some pretty specific information is provided on patients, e.g., 'French male tourist' with a date of the case report. I'm wondering if these individuals are consented to participate in the research and be effectively identified. This seems pretty specific. There is no information about consent.

Results/Phylogenetic analyses – the phylogenetic results are difficult to read. The image is low resolution, so if I zoom in on the pdf, it just gets blurry. There are a different nomenclatures being thrown about (A1, A1, GR, GH, etc.) and it is hard to find these groups on the tree and or their significance. There are no bootstrap values, so there is no confidence in the presented relationships. Again, if the authors use iTOL, they can make a more effective presentation with better labels on the tree. I don't see the pink clade on the main tree at all.

In the text, everything is presented in terms of GISAID accessions, which is nice for track but not for presentation. It would be better if the authors relabeled things in the supplementary table with more accessible labels like Moroc1, Franc1, Spain1, etc., and then not list every isolate in a clade in the text. It would be even more informative if you included isolation date in the name, e.g., Moroc1_3_21. This would allow readers to immediately interpret the phylogeny. As it is, it is uninformative with respect to isolation date and difficult to identify isolates from different countries.

Then in the sequence variant network analysis, the nomenclature seems changed again. Now there are things called 'IPMx'. I don't understand why this network analysis is not done in the context of the other non-Moroccan sequences. Given the phylogenetic result that the Moroccan sequences don't form a cluster by themselves, but rather cluster with other sequences with distinct geographic origins, it makes no sense to analyze them by themselves here. You are just picking up variants that differentiate the geographic clades and assigning them (incorrectly) to the Moroccan sequences. It would be much more powerful to compare the Moroccan sequences to the others WITHIN EACH CLADE as this allows for effectively independent replicates of viral adaptation to Moroccan hosts. Then you can look for convergently evolved changes associated with the Moroccan isolates.

Tables 1, 2, & 3 should identify the nonsynonymous changes (e.g., show the amino acid replacement in the same way you show the nucleotide substitution).

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Partly

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phylogenetics, population genetics, molecular evolution, bioinformatics, infectious disease

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research