SARS-CoV-2 tracking in Tunisia through next-generation sequencing: lessons for the future

Ahmed Rebai¹ · Amal Souissi¹ · Nabil Abid² · Saber Masmoudi¹

Received: 8 February 2021 / Accepted: 10 March 2021 / Published online: 2 April 2021
© Springer Nature Switzerland AG 2021

Abstract

In this study, data available from GISAID on the whole-genome sequences of SARS-CoV-2 variants circulating in Tunisia were analyzed, and the prevalences of those variants in Tunisia were compared to their prevalences in other North African countries and around the world. Our results show new mutations and different prevalences of some lineages. In particular, new sets of mutations were identified in the spike protein of the virus during the analysis of 85 Tunisian samples, and the lineage B.1.160 was found to be the most prevalent (18%) lineage in Tunisia. The prevalence of this lineage in Tunisia was significantly higher than its prevalence worldwide and in samples from neighboring countries (3%). This preliminary study shows the importance of tracking virus variants by next-generation sequencing in order to assess the dynamics of the COVID-19 pandemic and the impact of vaccination on the evolution of the virus.

Keywords COVID-19 · SARS-CoV-2 · Genome · Lineage · Coronavirus

Introduction

The outbreak of COVID-19 caused by the emergence of SARS-CoV-2 in December 2019 in China was declared a pandemic by the World Health Organization (WHO) on March 12, 2020. The rapid evolution of the pandemic as the viral infection has spread across the globe for over a year, resulting in over 100 million cases and 2.2 million deaths (as of January 27, 2021), has highlighted the harsh reality of the fragility of healthcare systems worldwide when facing such highly contagious and pathogenic viral infections.

Ever since the first cases were reported in China on December 12, 2019 (Wu et al. 2020), next-generation sequencing (NGS) technologies have been used to identify the causal pathogen, to understand its contagiousness and pathogenic features, and to track the evolution of SARS-CoV-2.

Since the publication of the first reference viral genome (NCBI accession number NC_045512.2), there has been a steady increase in the number of genome sequences published in public databases. More than two thousand variants/lineages are currently available in the GISAID database (<http://www.gisaid.org>), including recent reports of new variants from the UK, South Africa, and Brazil that are more contagious.

In Tunisia, the first wave of infection took place during March–May 2020, resulting in a thousand cases by June. Whole-genome sequencing has been performed at a limited scale, and an initial set of eight complete viral genome sequences were published in April 2020. The first local sequencing of SARS-CoV-2 genomes in Tunisia (using the single Illumina MiSeq available at the Center of Biotechnology of Sfax) was completed in August 11, 2020, and the resulting seven genomes were made available through the NCBI GeneBank (<http://www.ncbi.nlm.nih.gov>) on September 3, 2020 within the scope of a national project, ADAGE (Decision System Based on Genome Analyses in

Communicated by Mohamed Ksibi, Co-Editor in Chief.

This study was carried out on behalf of the ADAGE consortium, which includes 40 researchers from the Center of Biotechnology of Sfax, the University of Sfax, the University of Monastir, the University Hospital Hédi Chacker of Sfax, the Military Hospital of Tunis, and Dacima Consulting.

✉ Ahmed Rebai
ahmed.rebai@cbs.rnrt.tn

¹ Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax, P. O. Box 1177, 3018 Sfax, Tunisia

² Laboratory of Transmissible Diseases and Biological Active Substances, Faculty of Pharmacy, University of Monastir, 5000 Monastir, Tunisia

COVID-19 Tunisian Patients). The ADAGE project aims to sequence about 30 viral genomes and 100 human genomes collected from confirmed cases in Tunisia in order to better understand the role of the host–pathogen interaction in COVID-19 severity and obtain information that could lead to personalized healthcare for patients. The results of the project will integrate data on the virus and patient genomes as well as clinical data (e.g., severity and chronic disease) into an expert system through the use of machine learning algorithms. The aim of this is to establish optimized care for each COVID-19 patient according to the predicted severity of the disease.

Currently, Tunisia has recorded over 200,000 positive cases (45,000 of which are still active) and 6200 deaths (as of January 27, 2021), meaning that, among African countries, it has the second highest number of deaths relative to its total population. Eighty-five complete viral genomes collected from Tunisian cases are currently available in the NCBI and/or GISAID databases, as are hundreds of thousands of genomes from samples around the world (nearly 700,000 complete genome sequences as of March 5, 2021).

In this editorial, we describe the current state of knowledge about the diversity of SARS-CoV-2 viral variants circulating in Tunisia and North Africa, and compare that diversity to the diversity of the viral variants worldwide. We also discuss the importance of tracking this viral variant diversity in the management of the pandemic.

Material and methods

Data on complete viral genome sequences deposited in GISAID (<http://gisaid.org>) were uploaded (on January 15, 2021) and analyzed to identify mutations in comparison with the reference sequence of the Wuhan variant (accession NC_045512.2). The lineages of the sequences were determined based on data in GISAID according to the nomenclature described at <https://cov-lineages.org/> (Rambaut et al. 2020). Only sequences with metadata were considered, and patients were classified into two groups: symptomatic (mild or severe) and asymptomatic. After curating the data, we were able to collect both lineage and phenotype data for

5917 cases, among which 339 were asymptomatic and 5578 were symptomatic.

The frequencies of the lineages in different groups (countries or phenotype groups) were compared globally using the chi-square test in $r \times c$ contingency tables and pairwise by Fisher's exact test (FET) in 2×2 contingency tables. All statistical analyses were performed using appropriate functions in R (<http://www.r-project.org>).

Results and discussion: lessons learned from whole-genome sequencing in Tunisia

The first viral whole-genome sequences to be determined in Tunisia originated from a 31-year-old Tunisian soldier (who tested positive after returning from a mission in Morocco in March 2020) and a lady (59 years old) who was diagnosed at the Hopital Militaire de Tunis. Both were asymptomatic and carried two different viral lineages (B.1.255 and A, respectively) (Handrick et al. 2020).

Variant analyses

Twenty-two lineages were identified in the sample of 85 Tunisian cases. All but five Tunisian variants contained the D614G mutation in the spike (S) glycoprotein, either alone (28 strains; 33%) or with 1–4 other mutations. Among those five, four had an identical S glycoprotein to that of the reference variant, whereas one harbored a Y279N mutation. The latter was also found in one variant in GISAID. However, looking at the combinations of S glycoprotein mutations, it was clear that seven variants were unique and had not previously been reported in GISAID (Table 1). Note that one of these combinations (I233V/S477N/D614G) was recently reported in a sample from Mayotte (EPI_ISL_855374, submitted 01-21-2021). A high number of mutations occur in the spike protein because it is exposed on the surface of the viral particle and is the recognition site for both cell receptor and immune components. Indeed, mutations at a specific site in the S glycoprotein may hamper or enhance recognition events. Although the effects of some of these mutations have been explored, it is still unclear if the other mutations have

Table 1 List of unique mutations seen in Tunisian samples within the spike glycoprotein

Variant	Lineage	Reference in GISAID
R408K/D614G/Q675H	B.1.1.198	TUN_ADAGE_17-18245
P384L/E583D/D614G/M1237I	B.1.177.19	TUN_ADAGE_R-24804
A288T/Q314R/D614G	B.1.9	TUN/COV1339/2020
D111N/D614G/A892S	B.1.1	TUN202055304/2020
I233V/S477N/D614G	B.1.160	TUN202055319/2020
L5F/A222V/T572I/D614G	B.1.177	TUN202055400/2020
G261S/S477N/D614G/D627E/A1020S	B.1.160	hCoV-19/Tunisia/1268/2021

an impact on the contagiousness and pathogenic potential of SARS-CoV-2.

Other interesting features are the presence of two mutations that were each found in a single unique Tunisian variant (no other variants worldwide): S5L in ORF7b and E121Stop in ORF7a (from a sub-Saharan African patient). A variant with G8Stop and four variants with Q18stop in ORF8 in the Tunisian samples have been found in 118 and 1019 variants, respectively, worldwide (GISAID data on January 29, 2021). These two internal stop codons in ORF8 yield truncated proteins that are expected to be inactive. It is worth noting that the SARS-CoV virus of 2003 has two ORF8 proteins (ORF8a and ORF8b) (Yoshimoto 2020), and it was demonstrated that the protein ORF8b in SARS-CoV is involved in immune response regulation through binding to human interferon regulatory factor 3 (IRF3), a key transcriptional regulator of type I interferon (IFN)-dependent immune responses, which play a critical role in the innate immune response against both DNA and RNA viruses. Wong et al. (2018) have shown that the expression of ORF8b enables the virus to partially overcome the inhibitory action of IFN activation and thus achieve higher replication efficiencies in cells. The absence of ORF8 in SARS-CoV-2 may therefore result in a significantly reduced ability of the virus to replicate in host cells. Many recent studies (see Zinzula 2021 for a review) have reported several deletions and mutations in ORF8 that seem to be associated with milder symptoms and a better disease outcome. Note, however, that the UK variant (B.1.1.7) also contains a Q27Stop in ORF8 but does not show a reduced viral replication ability.

In summary, if we consider the combination of all protein mutations in the whole virus genome, 60 of the 85 Tunisian

samples/variants present 53 unique combinations, although some belong to the same lineage. The remaining 25 samples are reported in the GISAID database with different prevalence rates.

Comparison to North African countries

We were able to extract whole genome sequences and lineage data for 18 samples from Algeria and 121 samples from Morocco (GISAID on January 15, 2021). There were 39 lineages with at least one case in one of the populations, but only nine lineages that occurred at frequencies of more than 3% were considered for statistical analysis (Fig. 1).

The major lineage B.1 represented 15.3%, 50%, and 52.9% of the cases in Tunisia, Algeria, and Morocco, respectively. Interestingly, the two other major lineages in Tunisia (B.1.160 and B.1.177, which occurred at frequencies of 17.6% and 14.1%, respectively) were underrepresented in Morocco (1.6% and 2.5%, respectively) and absent in Algeria. This difference in lineage prevalence between Tunisia and Morocco (based on similar sample sizes) might be due to differences in the dynamic pattern of infection.

Although the three countries showed different variant distributions, they had three variants in common: B.1, which is a large European lineage that corresponds to the Italian outbreak; B.1.367, which is a French lineage; and B.1.1.119, which is a Northern Irish lineage.

Due to the limited number of Algerian samples, we focused on performing a statistical comparison between the Tunisian and Moroccan samples, and found that those two sets of samples had highly significantly different distributions of the nine lineages ($\chi^2 = 80.8$, $df = 8$, $p = 4 \times 10^{-14}$).

Fig. 1 Percent of cases (vertical axis) versus lineage for cases from three Maghreb countries (note that only the nine most prevalent lineages are represented), based on data available in GISAID as of January 15, 2021

Fig. 2 Global prevalences (vertical axis) of the 18 most prevalent lineages in the symptomatic and asymptomatic groups (based on data from GISAID as of January 15, 2021)

Specifically, the lineages B1.160, B1.177, and B1.177.14 were significantly more prevalent in Tunisia ($p=0.000009$, 0.001, and 0.0008, respectively), while the lineages B.1, B.1.1.119, and B.1.1.4 were more prevalent in Morocco ($p=0.0000001$, 0.001, and 0.014, respectively).

Based on the clinical data for 18 cases from Tunisia, we tried to correlate the virus variant/lineage to the severity of infection, but binary logistic regression of phenotype status (severe versus mild) adjusted for age and sex did not show any significant association with lineage. This is very likely due to the small sample size. However, we noticed that the most prevalent lineage in Tunisia (B.1.160) was present in three patients, all of whom showed a mild phenotype, whereas two patients with the lineages B.1.1.198 and two with B.1.1.1 showed severe symptoms.

Analysis of global data

GISAID metadata on infected cases were used to look for virus variants that showed different prevalences in the two severity groups. A total of 47 lineages in 5917 cases were examined, and only those that were present in more than five individuals

in at least one severity group were selected for further analyses (18 lineages; 288 asymptomatic and 4506 symptomatic cases). The distributions of those lineages differed highly significantly between the two severity groups ($\chi^2=607.5$, $df=17$, $p=4 \times 10^{-118}$), with five lineages overrepresented in asymptomatic cases and three showing greater prevalence in symptomatic cases (Fig. 2). Considering these lineages individually, highly significant differences in prevalence between the two groups were observed for lineages B, B.1.177, B.6, and B.1.1.229, which were significantly more prevalent in asymptomatic cases ($p=10^{-5}$, 10^{-11} , 6×10^{-9} and 5×10^{-48} , respectively). On the other hand, lineages B.1, B.1.1, and B.1.5 were more prevalent in symptomatic cases ($p=8 \times 10^{-15}$, 10^{-13} , and 10^{-11} , respectively). It is worth noting that lineage B.1.177 was the third most prevalent lineage in Tunisia (14% of cases) according to the data on the 85 cases available by the date of analysis.

Conclusion

The preliminary data presented here suggest that the pathogenicities of the variants circulating globally may differ between lineages, although the severity of the disease also depends on host genetic and nongenetic factors and on the complex interaction between host and pathogen. The emergence of new variants that are highly contagious and probably potentially pathogenic necessitates careful tracking. To achieve this goal, more effort should be directed into supporting the high-throughput sequencing of viral variants (mainly in highly affected middle- and low-income countries) to assess the circulating lineages, correlating these data with severity phenotypes, and tracking/assessing the efficiency of the protection provided by vaccines during and after the worldwide vaccination campaign.

Acknowledgements This work was funded by the Ministry of Higher Education and Scientific Research, Tunisia under the project ADAGE (PRFCOVID19-GP2), within the national strategy of facing the COVID-19 pandemic.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

- Handrick S, Bestehorn-Willmann M, Eckstein S et al (2020) Whole genome sequencing and phylogenetic classification of Tunisian SARS-CoV-2 strains from patients of the Military Hospital in Tunis. *Virus Genes* 56:767–771
- Rambaut A, Holmes EC, O'Toole Á et al (2020) A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 5:1403–1407
- Wong HH, Fung TS, Shouguo Fang S et al (2018) Accessory proteins 8b and 8ab of severe acute respiratory syndrome coronavirus suppress the interferon signaling pathway by mediating ubiquitin-dependent rapid degradation of interferon regulatory factor 3. *Virology* 515:165–175
- Wu F, Zhao S, Yu B et al (2020) A new coronavirus associated with human respiratory disease in China. *Nature* 579:265–269
- Yoshimoto FK (2020) The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *Protein J* 39:198–216
- Zinzula L (2021) Lost in deletion: the enigmatic ORF8 protein of SARS-CoV-2. *Biochem Biophys Res Comm* 538:116–124. <https://doi.org/10.1016/j.bbrc.2020.10.045>