Overview of genomic surveillance related to Severe Acute Respiratory Syndrom Coronavirus 2 (SARS- CoV-2)

Hanane BAHOUQ[1,2*], *Madiha* BAHOUQ[3], and *Abdelmajid* SOULAYMANI[2]

[1]Hospital of specialties Tangiers, Morocco.
[2]Laboratory of Genetic and Biometry, Faculty of Sciences, University Ibn Tofail, Kenitra, Morocco.
[3]Laboratory of Botanic, Biotechnologies and plants Protection, Faculty of Sciences, University Ibn Tofail, Kenitra, Morocco.

Abstract. Since the start of the Severe Acute Respiratory Syndrom Coronavirus 2 (SARS-CoV-2) pandemic, several thousand of variants circulated and others are emerging. Therefore, genomic surveillance is crucial, which aims to detect the emergence of new variants, in particular Variants of Concern (VOC) and to assess the impact of priority mutations on the transmissibility and lethality of the virus, the performance of viral diagnostic methods and vaccine efficiency. An overview of available papers was performed to understand conduct, tools and utility of genomic sequencing and surveillance related to Covid-19 disease. We also report the experience of Morocco in this filed through available data. A national SARS-Cov-2 genomic consortium has been established in order to continuously inform the health authorities of the genetic evolution of circulating strains in Morocco. Genomic sequencing shows that Moroccan genomes spread did not show a predominant SARS-CoV-2 lineage. Genomes are dispersed across the evolutionary tree of SARS-CoV-2 and held between 4 and 16 mutations. As the pandemic ongoing, continuous genomic surveillance and regular sequencing are fundamental to understand the spread of SARS-CoV-2, to rapidly identify potential global transmission networks and to consolidate response strategies especially targeted Covid-19 vaccination.

1. Introduction

Several coronavirus are already known to be able to infect humans. Severe Acute Respiratory Syndrom Coronavirus 2 (SARS- CoV-2), a lineage-b beta-coronavirus belonging to the coronaviridae family is the seventh coronavirus pathogenic to humans responsible for the COronaVIrus Disease 2019 (Covid-19). Covid-19 causes pneumonia and severe acute respiratory syndrome due to the high level of inflammatory response [1]. It was first identified in Wuhan, China, in early December 2019. Widespread cases were reported in other countries subsequently. In January, 2020, the World Health Organization (WHO) declared Covid-19 a Public Health Emergency of International Concern. The pandemic has infected more than 168,017,381 people and caused more than 3,488,273 deaths globally as of May 25 2021 (https:/www.worlmeters.info/coronavirus/). In Morocco, the first Covid-19 case was declared in March, 2020 with 517,113 confirmed cases and 9,126 deaths until May, 2020.

SARS-CoV-2 is a RiboNucleic Acid (RNA) enveloped virus. Its genome has a size of 30 kilobases approximatively, coding for 15 genes, 4 of which correspond to structural proteins: surface protein (Spike or S protein), envelope protein (E), matrix protein (M) and nucleocapsid protein (N) with accessory proteins, encoded by ORF3a, ORF6, ORF7a, ORF7b and ORF8 genes [2]. By the Receptor Binding Domain (RBD), the most variable part of the coronavirus genome, Spike protein is responsible of binding the Angiotensin-Converting Enzyme2 (ACE2) cell surface receptor and inducing cell entry during infection [3]. Despite SARS-CoV-2 lower mutation rate than most RNA viruses, mutations generally accumulate inducing genomic diversity. This purchased genetic heterogeneity procures viral adaptation to different hosts and environments. In hosts, genomic diversity, most often is associated with disease progression, drug resistance and vaccination issues. Since the beginning of the pandemic, several mutations of SARS-CoV-2 have been reported in the literature, most often, due to nucleotide substitution but gene deletions are also described. Consequently, genome sequencing of SARS-CoV-2 became a Public Health Priority. It aims to enable genomic epidemiology investigations into Covid-19 disease origins and spread,

*Corresponding author: hananebahouq@yahoo.fr

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

to contribute to a better understanding of viral pathogenesis and virulence and to provide support for targeted vaccines. As 26 May 2021, 1,732,197 SARS-CoV-2 genomic sequences have been shared via the Global Initiative on Sharing All Influenza Data (GISAID) database [4, 5]. In this global health crisis, to understand viral transmission and evolution and to inform public health responses and vaccine development, national and international consortiums were established. Open AI Consortium (COAI) was launched in order to advance collaborative research and accelerate clinical development of vaccines and treatments for patients infected with Covid-19 and to share findings with the global medical and scientific community. In parallel, Many Covid-19 Genomics Consortium were created to conduct rapid whole – genome sequencing to generate genetic knowledge of SARS-CoV-2 behaves and spreads and to enable the tracking and analysis of viral variants.

As other countries, in February 2021, the Moroccan health ministry has get up a consortium of laboratories with sequencing platform as a part of its strategy for genomic monitoring of the disease. This consortium is composed of the Reference Laboratory for Influenza and respiratory Viruses of the National Institute of hygiene, the Medical Biotechnology Laboratory of Faculty of Medicine and Pharmacy and the Functional Genomic Platform of the National Scientific Research as well as to the Institute Pasteur in Casablanca. The main mission of this laboratories network is to identify SARS-CoV-2 variants and to characterize them by genomic sequencing.

We chose to perform an overview as an appropriate approach to understand conduct, tools and utility of genomic sequencing and surveillance related to Covid-19 disease. Furthermore, we summarize how genomic sequencing and surveillance have supported the identification of new SARS-CoV-2 variants and mutations. Also, we report the available data of the Moroccan experience in this field.

2. Phylogenetic network analysis

In conducting genomic analysis, after specimen collection, DNA synthesis, genome viral amplification and next generation sequencing, genomes are mapped to the reference sequence Wuhan-Hu-1/2019 with Variant Call Forma (VCF). Then, a phylogenetic analysis is performed to construct the phylogenetic tree premising genomic comparison and analysis via the reference strain repositories such as GenBank and GISAID and to identify mutations [4, 6]. To date, the globally circulating viruses have been classified into six major clades denoted as S, L, V, G, GH, GR and GRY [5].

As the Covid-19 outbreak continues to evolve and scientific evidence expands rapidly, the information provided in this paper is only current as the date of elaborating this work.

2.1. Bioinformatics tools for genomic analysis

Many tools are available for each component step, from quality control of the genomic sequence data to viral genomic verification [7, 8]. Several bioinformatics tools have been developed for the detection and genomes SARS-CoV-2 sequencing (Covidex for SARS-CoV-2 genomes subtypification, CoV-GLUE for tracking SARS-CoV-2 genome accumulating changes, PoSeiDon for detection of positive selection in protein-coding genes, etc.) [7, 8]. The full-length genomic sequences and protein-coding sequences (CDSs) of SARS-CoV-2, SARS-CoV, MERS-CoV (Middle East respiratory syndrome) and bat coronaviruses are integrated into NCBI (National Center for Biotechnology Information), Severe acute respiratory syndrome coronavirus 2 data hub (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) and GenBank® database (http://www.ncbi.nlm.nih.gov/genbank/). GenBank ®, as a part of the International Nucleotide Sequence Database Collaboration, which comprises also the DNA DataBank of Japan (DDBJ) and the European Nucleotide Archive (ENA), is a genetic sequence database of all publicly available DNA sequences. Several bioinformatics tools accelerating SARS-CoV-2 research were developed with free use and online availability [7, 8].

Researchers, regularly deposit datasets into public databases such as GISAID for consensus sequence as the standard database for sharing of SARS-CoV-2 data internationally with no imposed limitations on the sharing and the genomic sequences use [5]. GISAID deposited the largest number of SARS-CoV-2 genome sequences.

The GISAID nomenclature system is based on marker mutations within eight high-level phylogenetic groupings or 8 Clades from the early split of S (pink) and to the further evolution of L (light gray) into V (magenta) and G (yellow), into GH (green), GR (red), GV and GRY [5]. The list of the marker variants is as follows:

S: C8782T,T28144C includes NS8-L84S

L: C241,C3037,A23403,C8782,G11083,G26144,T28144 (early clade markers in WIV04-reference sequence)

V: G11083T,G26144T NSP6-L37F + NS3-G251V

G: C241T,C3037T,A23403G includes S-D614G

GH: C241T,C3037T,A23403G,G25563T includes S-D614G + NS3-Q57H

GR: C241T,C3037T,A23403G,G28882A includes S-D614G + N-G204R

GV: C241T,C3037T,A23403G,C22227T includes S-D614G + S-A222V

GRY: C241T,C3037T,21765-21770del,21991 21993del,A23063T,A23403G,G28882A includes S-H69del, S-V70del, S-Y144del, S-N501Y + S-D614G + N-G204R

The definition of more detailed lineages is adopted through the Phylogenetic Assignment of Named Global Outbreak LINeages (PANGOLIN) tool allowed by Rambaut et al [5, 6, 9]. Lineages assigned by this software tool, are characterized by a combination of genetic and epidemiological support by describing a lineage as a cluster of sequences seen in a geographically distinct region with evidence of local ongoing

*Corresponding author: hananebahouq@yahoo.fr

transmission [5, 6, 9]. Pangolin lineage facilitate getting useful information of viral genome sequencing in real-time and can assist in identifying new introductions and in tracking the SARS-CoV-2 spread [6].

An additional classification effort has been provided by Hodcroft et al [10], which supports the PANGOLIN lineage nomenclature system and allows comparison to the SARS-CoV-2 reference sequence, assigning sequences to Clades and define where they fall on a the SARS-CoV-2 phylogenetic tree. Nextclade uses a Year-Letter nomenclature with a capital letter starting for the year when clade emerged. Once the frequency of a clade exceeds 20% for more than 2 months in a representative global sample, new major clades are used. Currently, clades 19A, 19B, 20A, 20B, 20C and 20I are named [10, 11].

In Morocco, the report of Badaoui [12] shows that Moroccan genomes are dispersed across the evolutionary tree of SARS-CoV-2. Viral strains are not only from Belgium, Spain and France but also from USA and Vietnam. Nine viruses from Clade 20A, 9 from 20B and 2 from 20C were included with no predominant SARS-CoV-2. The virus circulated on February 2020 before the official discovery of the first case in March [12].

2.2. SARS-CoV-2 variants and mutations

Since the start of the pandemic, thousands of mutations (amino acid replacement, insertion and deletion), which in turn have given rise to thousands of variants, were screened [13]. By convention an amino acid change is written N501Y to denote the wildtype (N, asparagine) and replacement amino acid (Y, tyrosine) at site 501 in the amino acid sequence.

In general, non-synonymous mutation, subject of natural selections, is a nucleotide mutation that alters the amino acid sequence of the spike protein which differs from synonymous substitution by silent mutation, without amino acid sequences alteration.

On 25 February 2021, WHO released a document outlining working definitions of variants of concern (VOC) and variants of interest (VOI) [14].

A VOI is defined as an isolate of SARS-CoV-2 with genotypic and/or phenotypic changes compared to the reference genome that have been associated with changes to receptor binding, reduced neutralization by antibodies generated against previous infection or vaccination, reduced efficacy of treatments, potential diagnostic impact, or predicted increase in transmissibility or disease severity.

A VOC is defined as a VOI which has an evidence of an increase in transmissibility, virulence and/or is not being controlled effectively by current public health measures.

On 31 May 2021, the WHO has assigned simple, easy to say and remember labels for key variants of SARS-CoV-2, using letters of the Greek alphabet. These labels do not replace existing scientific names (assigned by GISAID, Nextstrain and Pango), which convey important scientific information and will continue to be used in research [14].

2.2.1. Variants of Concern

To date, four variants of SARS-CoV-2 (B.1.1.7, B.1.351, P.1 and B.1.617.2) are subject of enhanced surveillance due to their considerable transmissibility and virulence.

In the United Kingdom (UK), the B.1.1.7 (Alpha), VOC-202012/01, 20B/501Y.V1 variant was the first variant identified as a VOC by the COVID-19 Genomics UK Consortium (GOC-UK) in November 2020 [15] and, currently, the most highly sequenced and well-characterized VOC. B.1.1.7 is known to have increased levels of transmissibility (40 and 70%) [16].

In South Africa, the B.1.351 (Beta), 501Y.V2; VOC 20C/501Y.V2 variant was identified by the Network for Genomic Surveillance in South Africa (NGS-SA) in December 2020 [17]. It has been shown to have increased transmissibility and to reduce the efficacy of some vaccines [18].

The Brazilian variant P.1 (Gamma), 501Y.V3 or B.1.1.28.1 VOC, was reported by Japan in December 2020 after detection in four travellers who had returned from Brazil [19]. Due to the presence of spike mutations (also found in the B.1.351 variant): N501Y and K417N/T (increase virus binding affinity to the ACE2 receptor on human cells and fast lineages growing with possible resistance to some antibodies), E484K (leads to escape from immune response); P.1 variant is flagged to be of concern.

The Indian variant B.1.617.2 (Delta) has emerged in India in December 2020 [20, 21] and was declared VOC by the UK in 7 May 2021. This variant is defined by four mutations in the Spike protein: E484Q, L452R (linked to increased transmissibility and virulence and immune protection evasion specifically targeting the spike RBD) and P681R (may increase the infectivity of the virus by facilitating cleavage site between S1/S2).

2.2.2. Variants of interest

For these seven variants, genomic and epidemiological evidence is available that could imply a significant impact on the epidemiological situation by significant transmissibility, severity and/or immunity [22]. We adopt the European classification of the European Centre for Disease Prevention and Control (ECDC) for VOI variants [22].

The B.1.525 (Eta) variant has emerged in Nigeria and UK in December 2020. It is a variant under investigation with still unknown infectivity level. B.1.525 is defined by 3 mutations: E484K, Q677H and F888L. The mutation of B.1.525 makes it similar to B.1.1.7 variant and may increase transmissibility, virulence, and immune escape [22].

The American variant, B.1.427/B.1.429 (Epsilon) has emerged in September 2021 with L452R and D614G (allows the virus rapidly replace strains without mutation, increase infectiousness with no reduced vaccine effectiveness) mutations [22].

The Indian variants B.1.617.1/B.617.3 (Kappa) have emerged in September 2021 and February 2021 respectively with L452R, E484Q, D614G and P681R mutations [22].

*Corresponding author: hananebahouq@yahoo.fr

The unclear variant origin B.1.620, has emerged in February 2021 with S477N (causes tight attachment to the ACE2), E484K, D614G and P681H mutations [22].

The Colombian variant B.1.621 has emerged in January 2021 with R346K, E484K, N501Y, D614G and P681H mutations [22].

The Philippian variant P.3 (Theta) has emerged in September 2020 with E484K, N501Y, D614G and P681H mutations [22].

The French variant B.1.616 reported in February 2021 with V483A, D614G, H655Y, and G669S mutations [22].

2.2.3. Variants under monitoring

Additional variants of SARS-CoV-2 have been detected which they could have properties similar to those of a VOC, but the evidence is weak or has not yet been scientifically assessed.

In our context, Badaoui [12] and Laamrati [23], reported that the virus genomes from Moroccan patients retain between 4 and 16 mutations relative to the common Wuhan-Hu-1/2019' ancestor. Most frequent non-synonymous mutations in SARS-CoV-2 isolates from Moroccan patients were nsp 12, P323L, D614G, R203K and G204R [12]. D614G nonsynonymous mutation, associated with the emergence of clade A2 and known as the most prevalent variant worldwide was found [23]. This mutation was already associated with the observed transmission increase in the United States.

In recent studies, others mutations were revealed. Twelve mutations in genome which belongs to clade 20A, (A2568T, C3037T, C5884T, C8169T, C9907G, C14408T, C17104T, A20268G, G21795T, A23403G, G25563T, and G29734C) were reported by Rfaki [24] and 34 variants assigned to the B.1.1.7 lineage, by Ouadghiri including N501Y mutation [25].

3. Conclusion

As the pandemic ongoing, continuous genomic surveillance and regular sequencing are fundamental to understand the spread of SARS-CoV-2 in different regions, to rapidly identify potential global transmission networks and to consolidate response strategies. Bioinformatics resources in sequence and phylogenetic alignment, tree visualization and genomic analysis are essentials. However, the increase in the amount of SARS-CoV-2 genome sequence data available represents serious challenges for data storage and analysis. National and international improvement of genomic surveillance tools and resources must be required to resolve this problem. Therefore, genome sequences updating in real time is crucial for tracking rapidly the genetic evolution of SARS-CoV-2 and the diffusion of emerging clades in order to develop the appropriate health strategies against circulating variants as well as new emerging VOC, especially in terms of Targeted Covid-19 Vaccination.

Ethical Approval and Consent to participate

Not applicable.

Funding

The authors declare that no funding was received for the present study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

4. References

1. K.G. Andersen, A. Rambaut, W.I. Lipkin, E. C. Holmes, R. F. Garry, Nat. Med **26,** 450–452 (2020)
2. R.A. Khailany, M. Safdar, M. Ozaslan, Gene Rep **19** (2020)
3. A. C. Walls, Y. J. Park, M. A. Wall, A. T. McGuire, D. Veesler Cell **181**, 281-292 (2020).
4. Global initiative on sharing all influenza data. GISAID, available online from: https://www.gisaid.org/ (Accessed 2021 May 2021).
5. GISAID-Clade and Lineage Nomenclature Aids in Genomic Epidemiology of Active hCoV-19 Viruses, available online from: https://www.gisaid.org/references/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomicepidemiology- of-active-hcov-19-viruses/ (Accessed 2021 May 14).
6. The Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN), available online from: https://pangolin.cog-uk.io/ and as command line tool via GitHub: (https://github.com/hCoV-2019/pangolin/ (Accessed 2021 May 14)
7. H. Tao, L. Juan, Z. Hong, L. Cixiu, H. Edward C, W. Shi, Brief. Bioinformatics **22**, 631-641 (2021)
8. M. Ray, M. N. Sable, S. Sarkar, V. Hallur, Meta Gene **27** (2021)
9. A. Rambaut, E.C. Holmes, Á. O'Toole, V. Hill, J.T. McCrone, C. Ruis, L. du Plessis, O.G. Pybus, Nat. Microbiol **5**, 1403–1407 (2020)
10. E. B. Hodcroft, J. Hadfield, R. A. Neher, T. Bedford. Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org, available online from: https://nextstrain.org//blog/2020-06-02-SARSCoV2-clade-naming (Accessed 2021 May 14)
11. Updated Nextstain SARS-CoV-2 Clade Naming Strategy-SARS-CoV-2 Coronavirus/nCoV-2019 Genomic Epidemiology, available online from: https://virological.org/t/updated-nextstain-sars-cov-2-clade-naming-strategy/581 (Accessed on 15 May2021).
12. B. Badaoui, K. Sadki, C. Talbi, D. Salah, L. Tazi, Biosaf Health **3,** 124–127 (2021)
13. CoV-GLUE enabled by data from GISAID 2021, available online from: http://cov-glue.cvr.gla.ac.uk/#/home (Accessed 2021 May 14)
14. Weekly epidemiological update. World Health Organization, available online from: https://www.who.int/publications/m/item/covid-19-weekly-epidemiological-update.
15. The COVID-19 UK Genomics Consortium COG-UK. (2021) available online from: https://www.cogconsortium.uk/ (Accessed 2021 May 14)
16. E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, R. W. Hinsley, D.J. Laydon, G. Dabrera, A. O'Toole, R. Amato, M. Ragonnet-Cronin, I. Harrison, B. Jackson, C. V. Ariani, O. Boyd, N.J. Loman, J.T. McCrone, S. Gonçalve, N.M. Ferguson, Nature, 1–17 (2021)

*Corresponding author: hananebahouq@yahoo.fr

17. Network for Genomic Surveillance in South Africa. KRISP, available online from: https://www.krisp.org.za/ngs-sa/ngasa_network_members_nhls_nicd_uct_ukzn_h3abionet_and_others/ (Accessed 2021 May 14).
18. S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, AL. Koen, L. Fairlie, SD. Padayachee, K. Dheda, SL. Barnabas, QE. Bhorat, C. Briner, G. Kwatra, K. Ahmed, P. Aley, S. Bhikha, J.N. Bhiman, A.E. Bhorat, J. du Plessis, A. Esmail, M. Groenewald, E. Horne, S.H. Hwa, A. Jose, T. Lambe, M. Laubscher, M. Malahleha, M. Masenya, M. Masilela, S. McKenzie, K. Molapo, A. Moultrie, S. Oelofse, F. Patel, S. Pillay, S. Rhead, H. Rodel, L. Rossouw, C. Taoushanis, H. Tegally, A. Thombrayil, S. van Eck, C.K. Wibmer, N.M. Durham, E.J. Kelly, T.L. Villafana, S. Gilbert, AJ. Pollard, T. de Oliveira, P.L. Moore, A. Sigal, A. Izu, N Engl J Med **384**, 1885-1898 (2021)
19. Brief report: New Variant Strain of SARS-CoV-2 Identified in Travelers from Brazil. National Institute of Infectious Diseases, Japan, available online from: https://www.niid.go.jp/niid/en/2019-ncov-e/10108-covid19-33-en.html (Accessed 2021 May 14)
20. PANGO lineages. SARS-CoV-2 lineages, available online from: https://cov. lineages.org/global_report.html (Accessed 2021 May 14)
21. Global Virus Network. Covid-19 Variants and Vaccines, available online from: https://gvn.org/home/ (Accessed 2021 May 14)
22. European Centre for Disease Prevention and Control (ECDC), SARS-CoV-2 variants of interest available online from: https://www.ecdc.europa.eu/en/covid-19 Accessed 2021 May 24)
23. M. Laamarti, M.W. Chemao-Elfihri, S. Kartti, R. Laamarti, L. Allam, M. Ouadghiri, I. Smyej, J. Rahoui, H. Benrahma, I. Diawara, T. Alouane, A. Essabbar, S. Siah, M. Karra, N. El Hafidi, R. El Jaoudi, L. Sbabou, C. Nejjari, S. Amzazi, R. Mentag, L. Belyamani, A. Ibrahimi, Microbiol Resour Announc **9** (2020)
24. A. Rfaki, N. Touil, M. Hemlali, S. Alaoui Amine, M. Melloul, M.A. El Alaoui, H. Elannaz, A.I. Lahlou, M. Elouennass, K. Ennibi, E. El Fahime, Microbiol Resour Announc **10** (2021).
25. M. Ouadghiri, T. Aanniz, A. Essabbar, M. Seffar, H. Kabbaj, G. El Amin, A. Zouaki, S. Amzazi, L. Belyamani, A. Ibrahimi, Microbiol Resour Announc **10** (2021).

*Corresponding author: hananebahouq@yahoo.fr