Introduction

Even with the extraordinary progress in biomedical sciences, the knowledge and understanding of zoonoses and pathogens diversity is still very limited (1). RNA viruses are proficient at switching to novel host species due to their fast mutation rates. Implicit in this assumption is the need to evolve adaptations in the new host species to exploit their cells efficiently. However, SARS-CoV-2 has required no significant adaptation to humans since the pandemic began (2). Due to their error prone polymerase/reverse-transcriptase (= 10-4/site/replication cycle), RNA viruses have more genetically diversified populations, are very variable in sequence, and different isolates from the same virus usually have many single nucleotide polymorphisms. A typical SARS-CoV-2 virus accumulates two single-letter mutations per month in its genome, a rate of change about that of Influenza and one-quarter that of HIV (3). In this study we intend to characterize the genetic diversity of the complete SARS-CoV-2 genome.

Aim

➢ Characterization of the genetic diversity of SARS-CoV-2
➢ Identification of the most conserved genomic regions of SARS-CoV-2

Methodology

We used GISAID (gisaid.org) and COVID CG (covidcg.org) to retrieve the complete SARS-CoV-2 genome sequences of all variants. We then used an in-house python script to automatically obtain sequences of the SARS-CoV-2 reference genome. The sequences with unidentified nucleotides (n=1) were eliminated. An alignment of all sequences was performed to detect the intraspecific genetic diversity, using the software MUSCLE (4).

A sliding window analysis was made using the final SARS-CoV-2 complete genomes alignment. We used windows with length of 50, 100, 200, and 300 nucleotides (nt). Afterwards we calculated the percentage of identical sites along the alignment of the SARS-CoV-2 complete genomes sequences using a python algorithm developed by the team.

Results

Our results demonstrate clearly the parts of the genome that are most conserved (Table 1 and Figure 1). We detected the highest values of percentage of identical sites (PIS) between the 28000 and 30000 nucleotide position in the alignment (Table 1). These regions contain the nucleocapsid phosphoprotein, N protein, an RNA-binding protein critical for viral genome packaging and viral assembly, and the Open reading frame 10 (ORF10), a unique SARS-CoV-2 accessory protein, which contains eleven cytotoxic T lymphocyte (CTL). The highest values of PIS and the respective window length were annotated (Figure 1).

Table 1: Maximum values of percentage of identical sites for the SARS-CoV-2 alignment considering different windows length (50, 100, 200, and 300 nt).

<table>
<thead>
<tr>
<th>Window length</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>End</td>
<td>Start</td>
<td>End</td>
<td>Start</td>
</tr>
<tr>
<td>28354</td>
<td>28404</td>
<td>28345</td>
<td>28445</td>
<td>28346</td>
</tr>
<tr>
<td>28762</td>
<td>28812</td>
<td>28753</td>
<td>28853</td>
<td>28754</td>
</tr>
<tr>
<td>28780</td>
<td>28830</td>
<td>28771</td>
<td>28871</td>
<td>28772</td>
</tr>
<tr>
<td>28798</td>
<td>28848</td>
<td>28789</td>
<td>28889</td>
<td>28790</td>
</tr>
</tbody>
</table>

Figure 1: Line chart showing the percentage of identical sites variation along the SARS-CoV-2 genome considering different windows lengths (50, 100, 200, and 300 nt). Regions with highest values of PIS were annotated with red rectangles.

Conclusion

Our analysis suggests that several highly conserved regions are present in the SARS-CoV-2 genome. These regions can be used to improve the detection methods of the SARS-CoV-2 virus (e.g., design highly efficient primers).

Acknowledgements

This research was supported by national funds through FCT (Foundation for Science and Technology) within the scope of UIDB/04423/2020 and UIDP/04423/2020. J.C. also acknowledges the FCT funding for his research contract at CIIMAR, established under the transitional rule of Decree Law 57/2016, amended by Law 57/2017.

References


