 doi:10.20944/preprints202006.0250.v1

Evolution of nucleotide composition in the SARS-CoV-2 lineage: Implications for vaccine design

Sankar Subramanian

GeneCology Research Centre, School of Science and Engineering, The University of the Sunshine Coast, Moreton Bay, QLD 4502, Australia

Running head: Evolution of nucleotide composition in the SARS-CoV-2 lineage

Keywords: Nucleotide composition, mutational change, GC to AT, COVID-19, SARS-CoV-2, Vaccine design

Address for correspondence:

GeneCology Research Centre

School of Science and Engineering

University of the Sunshine Coast

90 Sippy Downs Drive

Sippy Downs QLD 4556

Australia

Phone: + 61-7-5430 2873

Fax: +61-7- 5430 2881

E-mail: ssankara@usc.edu.au

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

Abstract

The worldwide outbreak of a novel coronavirus, SARS-CoV-2 has caused a pandemic of respiratory disease. Due to this emergency, researchers around the globe have been investigating the evolution of the genome of SARS-CoV-2 in order to design vaccines. Here I examined the evolution of GC content of SARS-CoV-2 by comparing the genomes of the members of the group *Betacoronavirus*. The results of this investigation revealed a highly significant positive correlation between the GC contents of betacoronaviruses and their divergence from SARS-CoV-2. The betacoronaviruses that are distantly related to SARS-CoV-2 have much higher GC contents than the latter. Conversely, the closely related ones have low GC contents, which are only slightly higher than that of SARS-CoV-2. This suggests a systematic reduction in the GC content in the SARS-CoV-2 lineage over time. The declining trend in this lineage predicts a much-reduced GC content in the coronaviruses that will descend/evolve from SARS-CoV-2 in the future. Due to the three consecutive outbreaks (MERS-CoV, SARS-CoV and SARS-CoV-2) caused by the members of the SARS-CoV-2, the scientific community is emphasizing the need for universal vaccines that are effective across many strains including those, that will inevitably emerge in the near future. The reduction in GC contents implies an increase in the rate of GC→AT mutations than that the mutational changes in the reverse direction. Therefore, understanding the evolution of base composition and mutational patterns of SARS-CoV-2 could be useful in designing broad-spectrum vaccines that could identify and neutralize the present and future strains of this virus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

Introduction

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), which belongs to the *Betacoronavirus* group, caused a pandemic by inflicting respiratory illness in human populations around the globe. The outbreak caused by SARS-CoV-2 resulted in >8 million cases and >400K deaths to date. A number of previous studies have characterized the genome of SARS-CoV-2 including its nucleotide composition. Although these studies compared the G+C content of this genome with those of the other members of the genus *Betacoronavirus* (1-5), the evolution of GC content in the SARS-CoV-2 lineage is unknown. Understanding the evolution of base composition in this virus has important implications in designing vaccines.

Materials and Methods

Genome data

From the data resource *GenBank* we obtained the complete genome sequences of seven viruses belonging to the subgenus *Sarbecovirus*: human SARS-CoV-2 (NC_045512), bat RaTG13 (MN996532), Pangolin-PCoV_MP789 (MT121216), Pangolin-PCoV_GX-P5E (MT040336), bat SL-CoVZC45 (MG772933), human SARS-CoV (NC_004718) and bat BM48-31 (NC_014470). We also obtained one genome each from the subgenera *Hibecovirus* (bat Zhejiang2013-NC_025217), *Nobecovirus* (bat HKU9-1-NC_009021) and *Merbecovirus* (human MERS-CoV-NC_038294) genomes. An *Embecovirus* (human HKU1-NC_006577) genome was also included for the purpose of using it as an outgroup. Using GenBank annotations, CDS regions were extracted for protein coding genes. Since all mutations at third codon positions of eight amino acids (Alanine, Arginine, Glycine, Leucine, Proline, Serine, Threonine, Valine) do not change the amino acids coded by the respective codons, these positions are called fourfold degenerate sites. We used them as the proxy for synonymous sites.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

Phylogenetic Analysis

We used the Maximum Likelihood method to infer the phylogenetic relationship among the viral genomes. This was accomplished using the *RaxML* program (13). The sophisticated General Time Reversible (GTR) method was used to model the evolution of nucleotides, as it addresses the differential rates of nucleotide changes including transition/transversional and base compositional biases. To accommodate rate variations among nucleotide sites, we used the gamma model of evolution. To test the strength of phylogenetic relationships we used the bootstrap resampling procedure with 500 replications.

To compute pairwise evolutionary divergence, we first estimated the shape of the rate variation among site (gamma) using the program MEGA (14). For this purpose, the Maximum Likelihood method was used. The HKY model of nucleotide evolution and the discrete gamma model with five categories were used to compute the gamma value. This analysis produced a gamma value of 0.54, which was used to estimate the pairwise evolutionary divergence using the Maximum Composite Likelihood method. G+C contents were estimated by counting of the G's and C's in the genome (or fourfold degenerate sites) and dividing this by the total number of nucleotides.

Results and Discussion

The GC content of the whole genome of SARS-CoV-2 is 37.97%. To understand how the observed GC content has evolved in the SARS-CoV-2 lineage, I collected complete genome sequences of all representative members of the subgenus *Sarbecovirus* including two human (SARS-CoV and SARS-CoV-2), three bat (RaTG13, SL-CoVZC45 and BM48-31) and two Pangolin (PCoV_MP789 and PCoV_GX-P5E) viruses. I also obtained the genomes of a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

Hibecovirus (Zhejiang2013), *Nobecovirus* (HKU9-1) and a *Merbecovirus* (MERS-CoV). In addition to the above, an *Embecovirus* (HKU1) was included for the purpose of using it as the outgroup. These genomes were used to construct a Maximum Likelihood phylogenetic tree which is shown in Figure 1A. The topology of this tree was very similar to those published by previous studies (6, 7). The GC contents of these genomes were estimated and are shown in Figure 1B. As it is clear from comparing Figures 1A and 1B, the GC contents gradually declines from MERS-CoV to SARS-CoV-2 viruses. To further explore this relationship, pair-wise distances between SARS-CoV-2 and the nine other viruses were calculated, and these estimates were plotted against the GC contents of the respective genomes. A highly significant positive correlation (Spearman $\rho = 0.9$, $P = 0.002$) between the pairwise evolutionary divergence and GC content was observed (Figure 2A). This suggests that the difference in the GC content of the virus that is closely related to SARS-CoV-2 is much smaller than those estimated using distantly related viruses. For instance, the evolutionary divergence between SARS-CoV-2 and the pangolin virus, PCoV_MP789 is 0.12 substitutions/site and the difference in their genomic GC contents is also very small 0.2%. In contrast, the divergence between SARS-CoV-2 and MERS-CoV viruses is 1.27 substitutions/site and the difference between their GC contents was 3.2%, which is 14 times higher than the difference observed for the pangolin virus.

Since ~98% of the SARS-CoV-2 virus is occupied by protein-coding genes the genome is under high purifying selection. Due to this reason the base composition evolution is also constrained by selection. Therefore, we used the nucleotides constituting the third positions of synonymous codons and estimated their GC contents. As expected the relationship between evolutionary divergence and GC content at synonymous positions was also highly significant ($\rho = 0.9$, $P = 0.002$) as shown in Figure 2B. However, the variation in the synonymous site

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

GC contents were much higher than that observed for the whole genomes. The GC contents of the genome range between 37.97% to 41.17% and hence the difference in the genomic GC contents is only 3.2%. On the contrary, GC contents of synonymous sites range between 20.21% to 28.65%, and therefore difference is 8.43%, which is 2.6 times higher than that observed for the genome.

Previous studies suggested that the synonymous site GC content is driven by codon usage bias (1, 3, 4). However, this is unlikely in the present case as a highly significant GC correlation was observed for the whole genomes as well. To further support this, we correlated the GC contents of nonsynonymous sites with evolutionary divergence and observed a highly significant relationship ($\rho = 0.98$, $P < 0.001$). Hence this suggest that the GC evolution is occurring throughout the genome including the constrained nonsynonymous sites and therefore it is unlikely to be driven by codon usage bias.

In this study we traced back the GC content of the SARS-CoV-2 lineage until the subgenera of *Merbecovirus*. However, the GC content of the outgroup taxon *Embecovirus* (genomic GC=32%) was found to be much smaller than that of *Merbecovirus* (genomic GC=41.6%). Hence it can be assumed that the ancestor of *Merbecovirus*, *Nobecovirus*, *Hibecovirus and Sarbecovirus* had the highest fraction of G+C nucleotides. This high GC content started to decline throughout the lineage leading to SARS-CoV-2, which contains the lowest GC content. Based on this declining trend shown in Figure 1 it can be predicted that any virus that derive from this lineage in the near future will more likely have a GC content that is much lower than that observed for SARS-CoV-2 (37.97%). Furthermore, similar prediction could be made if we compare only the human SARS like viruses.

 doi:10.20944/preprints202006.0250.v1

In the past two decades, three outbreaks were caused by the viruses (MERS-CoV, SARS-CoV and SARS-CoV-2) belonging to the same lineage (8). Hence future pandemic outbreaks appear to be highly likely to be caused by a novel SARS virus that descends from the SARS-CoV-2 lineage (9). The GC contents of the synonymous sites of these three coronaviruses (MERS-CoV, SARS-CoV and SARS-CoV-2) are 29%, 25% and 20% respectively. This suggests that there was a 14% reduction in the GC content of the SARS-CoV compared to MERS-CoV virus from which the former has evolved or derived (29% Vs 25%). Similarly, a 20% reduction in the GC content of the SARS-CoV-2 compared to that of SARS-CoV virus from which the former is descended. Hence any novel virus (SARS-CoV-X) that descends from SARS-CoV-2 lineage will potentially have a much-reduced GC content.

Due to the past outbreaks caused by the viruses belonging to the same lineage, recent studies emphasize the need for designing vaccines that are effective on not only the current SARS-CoV-2 strains but also the novel SARS viruses that eventually evolve from this lineage (9-12). As explained above, the results of the present study could be useful in predicting the GC content of future SARS viruses and hence help in designing universal vaccines. For example, reduction in GC content suggests a higher rate of GC→AT mutations. Hence, we can predict that the G and C nucleotides in genomic regions (eg. of the *S* gene) constituting the epitopes are more likely to be mutated to A or T compared to the mutations in the opposite direction. Therefore, knowledge about the most probable future mutation types is immensely useful in designing broad-spectrum epitopes that recognizes many strains of SARS-CoV-2 and its descendants emerging the imminent future.

Acknowledgement

This work was supported by a funding from the University of the Sunshine Coast (DVC-Research & Innovation grant).

References

1. Berkhout B, van Hemert F. 2015. On the biased nucleotide composition of the human coronavirus RNA genome. Virus Res 202:41-7.
2. Grigoriev A. 2004. Mutational patterns correlate with genome organization in SARS and other coronaviruses. Trends Genet 20:131-5.
3. Gu H, Chu DKW, Peiris M, Poon LLM. 2020. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. Virus Evol 6:veaa032.
4. Tort FL, Castells M, Cristina J. 2020. A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. Virus Res 283:197976.
5. Xia X. 2020. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Mol Biol Evol doi:10.1093/molbev/msaa094.
6. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565-574.
7. Tang T, Changcheng W, Xiang L, Yuhe S, Xinmin Y, Xinkai W, Yuange D, Hong Z, Yirong W, Zhaohui Q, Jie C, Jian L. 2020. On the origin and continuing evolution of SARS-CoV-2. National Science Review 0:1-12.

8. Tang D, Comish P, Kang R. 2020. The hallmarks of COVID-19 disease. PLoS Pathog 16:e1008536.

9. Giurgea LT, Han A, Memoli MJ. 2020. Universal coronavirus vaccines: the time to start is now. NPJ Vaccines 5:43.

10. Wang C, Li W, Drabek D, Okba NMA, van Haperen R, Osterhaus A, van Kuppeveld FJM, Haagmans BL, Grosveld F, Bosch BJ. 2020. A human monoclonal antibody blocking SARS-CoV-2 infection. Nat Commun 11:2251.

11. Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, Mok CKP, Wilson IA. 2020. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science 368:630-633.

12. Padron-Regalado E. 2020. Vaccines for SARS-CoV-2: Lessons from Other Coronavirus Strains. Infect Dis Ther doi:10.1007/s40121-020-00300-x:1-20.

13. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-3.

14. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547-1549.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

Figure Legends

Figure 1. (A) A Maximum Likelihood tree showing the phylogenetic relationship among *Sarbecovirus* (SARS-CoV-2, RaTG13, PCoV_MP789, PCoV_GX-P5E, SL-CoVZC45, SARS-CoV and BM48-31), *Hibecovirus* (Zhejiang2013), *Nobecovirus* (HKU9-1) and *Merbecovirus* (MERS-CoV) genomes. An *Embecovirus* (HKU1) genome was used as the outgroup. All nodes have at least 74% bootstrap (500 replications) support. **(B)** Genomic GC contents estimated for the 10 ingroup viruses are shown and the error bars show standard error of the mean.

Figure 2. Correlation between evolutionary divergence and GC contents of the Betacoronaviruses. Pairwise evolutionary divergences between SARS-CoV-2 and nine other viruses were estimated. **(A)** Whole genomes ($\rho = 0.9$, $P = 0.002$). **(B)** Synonymous positions ($\rho = 0.9$, $P = 0.002$). Best fitting regressing lines are shown.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

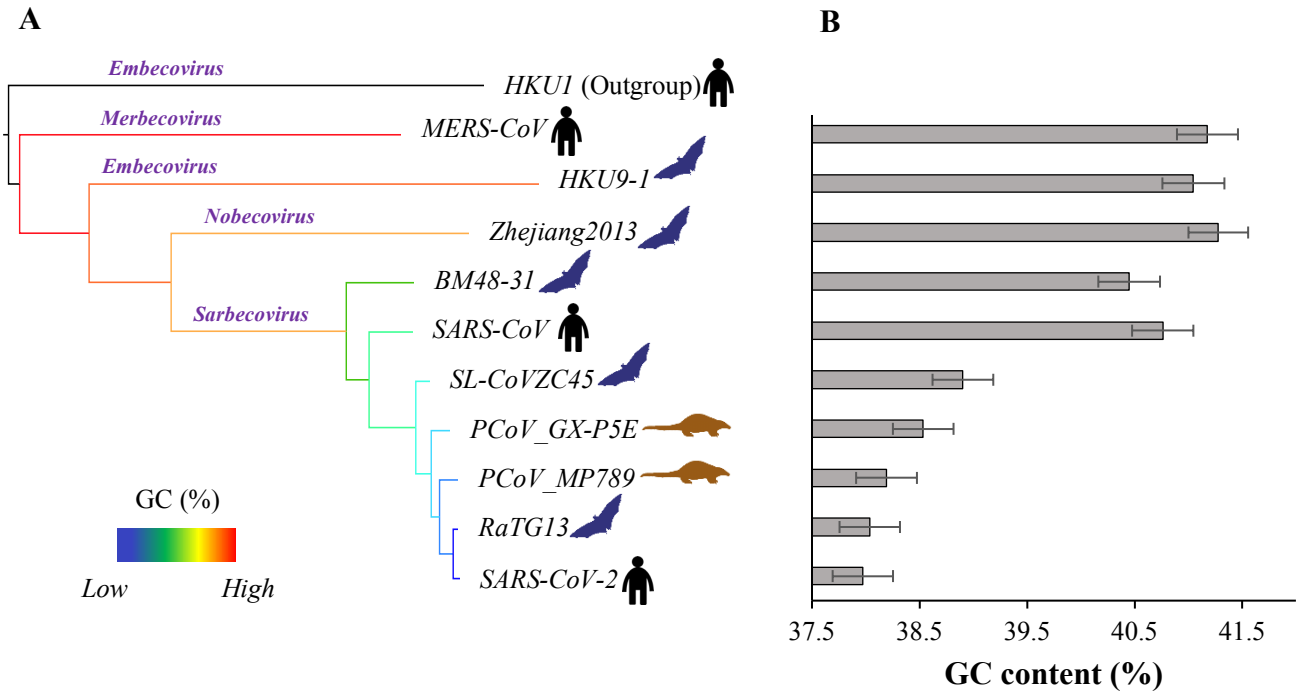

Figure 1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 June 2020 doi:10.20944/preprints202006.0250.v1

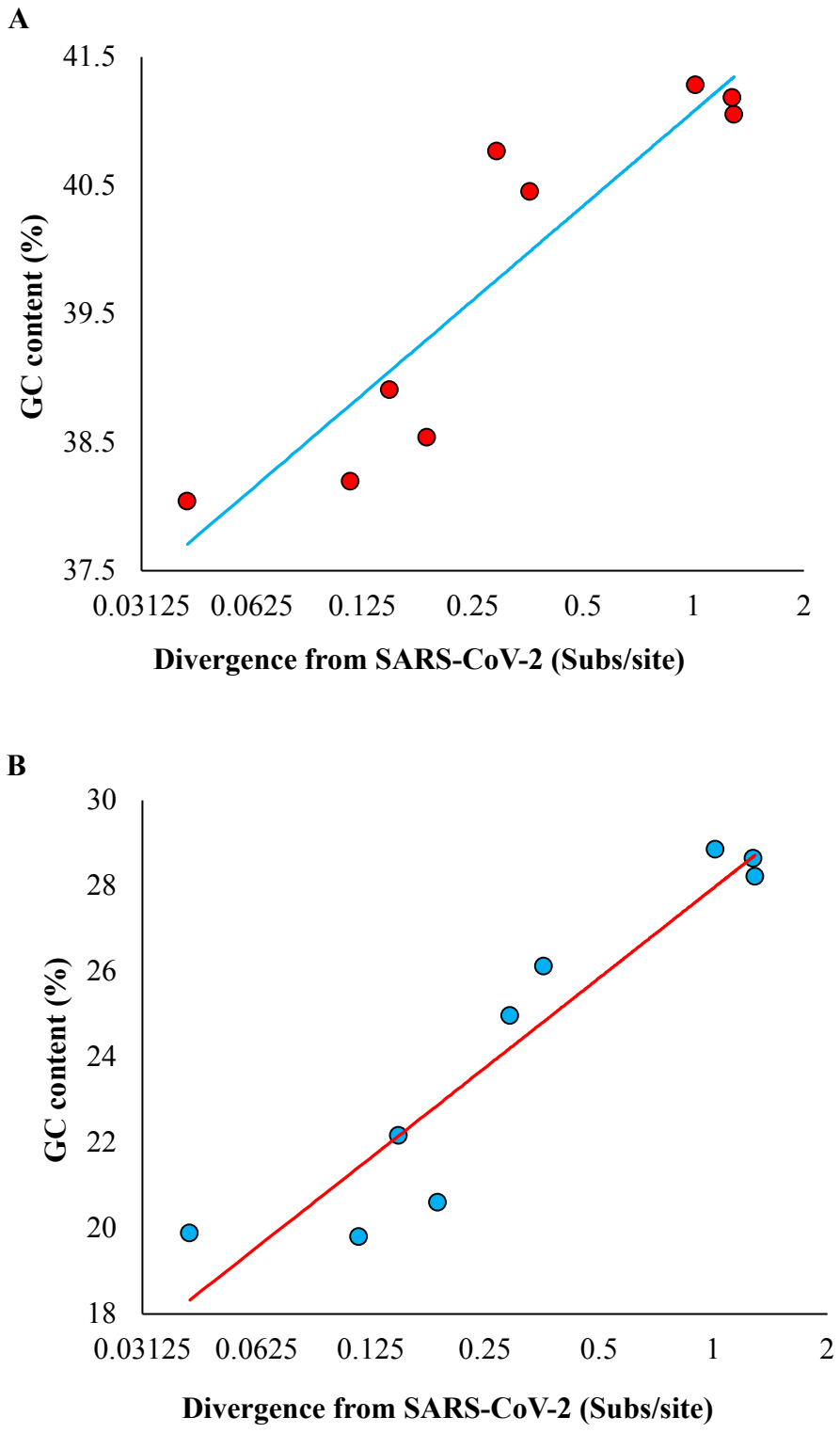

Figure 2