Genome analysis

Persistent minimal sequences of SARS-CoV-2

Diogo Pratas [1,2,3,*] **and Jorge M. Silva** [1,2]

[1] IEETA and [2] DETI, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
[3] Department of Virology, University of Helsinki, PL 21 (Haartmaninkatu 3) 00014, Helsinki, Finland

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused more than 14 million cases and more than half million deaths. Given the absence of implemented therapies, new analysis, diagnosis, and therapeutics are of great importance.
Results: Analysis of SARS-CoV-2 genomes from the current outbreak reveals the presence of short persistent DNA/RNA sequences that are absent from the human genome and transcriptome (PmRAWs). For the PmRAWs with length 12, only four exist at the same location in all SARS-CoV-2. At the gene level, we found one PmRAW of size 13 at the Spike glycoprotein coding sequence. This protein is fundamental for binding in human ACE2 and further use as an entry receptor to invade target cells. Applying protein structural prediction, we localized this PmRAW at the surface of the Spike protein, providing a potential targeted vector for diagnostics and therapeutics. Additionally, we show a new pattern of relative absent words (RAWs), characterized by the progressive increase of GC content (Guanine and Cytosine) according to the decrease of RAWs length, contrarily to the virus and host genome distributions. New analysis shows the same property during the Ebola virus outbreak. At a computational level, we improved the alignment-free method to identify pathogen-specific signatures in balance with GC measures and removed previous size limitations.
Availability and Implementation: `https://github.com/cobilab/eagle`.
Contact: pratas@ua.pt
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the past two decades, beta *coronaviruses* (CoV) have caused three zoonotic outbreaks, SARS-CoV in 2002-2003, MERS-CoV in 2012, and the newly emerged SARS-CoV-2 in late 2019 (Drosten *et al.*, 2003; Zaki *et al.*, 2012; Zhou *et al.*, 2020). SARS-CoV-2, initially referred to as nCoV-19, is a positive-sense single-strand RNA virus with a tracked origin to a food market in Wuhan, China, in December 2019 that can cause COVID-19 disease (Wu *et al.*, 2020). SARS-CoV-2 transmits by inhalation or contact with infected droplets having an incubation period from 2 to 14 days (Wang *et al.*, 2020).

The clinical features of COVID-19 are varied, common symptoms are fever, cough, shortness of breath, headache, while muscle pain, sputum production, diarrhea, and sore throat are less common (Chen *et al.*, 2020). The majority of cases result in mild symptoms and only a few progress to ARDS and multi-organ failure (Huang *et al.*, 2020; Lai *et al.*, 2020).

According to the World Health Organization (WHO), SARS-CoV-2 already exceeded the Ebolavirus outbreak, continuously escalating, with over 14 million cases and more than half million deaths confirmed in more than 180 countries, mainly in the USA, Brasil, India, Russia, and Peru.

Several therapeutic candidates are being rushed to face the pandemics, including successful in-vitro drugs, such as Remdesivir and Chloroquine, plasma transfusions and ACE2 based targets (Gurwitz, 2020; Zhang *et al.*, 2020; Nguyen *et al.*, 2020). Currently, the treatment is mainly supportive and symptomatic, where oxygen therapy represents the primary intervention for patients with severe infection. Therefore, new analysis with SARS-CoV-2 signatures may provide potential directions for the development of novel diagnostics and therapeutics.

Minimal Relative Absent Words (mRAW) are the shortest DNA (or RNA) sequences that exist in a pathogen and are absent from its host DNA, expressing pathogen-signatures with the potential to build fast diagnostic and targeted therapeutics (Silva *et al.*, 2015). mRAWs are a particular set of Minimal Absent Words (MAWs), a concept that has been studied since

1996 (Béal *et al.*, 1996; Crochemore *et al.*, 1998; Pinho *et al.*, 2009; Wu *et al.*, 2010; Chairungsee and Crochemore, 2012; Falda *et al.*, 2014; Garcia *et al.*, 2011; Herold *et al.*, 2008; Héliou *et al.*, 2017; Barton *et al.*, 2014; Vergni and Santoni, 2016; Crochemore *et al.*, 2020; Santoni and Vergni, 2020).

In this article, we introduce persistent mRAWs (PmRAWs), as fully conserved mRAWs across a gene or genomes in all the viruses. The methods combine PmRAWs with GC measures, alignments, and structural protein predictions. We localize specific signatures of the SARS-CoV-2 genomes and localize the PmRAWs in the predicted protein structure, using protein structural modeling followed by motif search. Additionally, we unveil a GC content pattern that characterizes mRAWs, which is present in both SARS-CoV-2 and Ebolavirus (EBOV) outbreaks.

2 Methods

2.1 Persistent minimal Relative Absent Words (PmRAWs)

Consider a set, X, constituted of n target sequences, $x_1, x_{\ldots}, x_n$, and a reference sequence, y, both drawn from the finite alphabet Θ. We say that β is a factor of x_i if x_i can be expressed as $x_i = u\beta v$, with uv denoting the concatenation between sequences u and v. We denote by $\mathcal{W}_k(x_i)$ the set of all k-size words (or factors) of x_i. Also, we represent the set of all k-size words *not in* x_i as $\overline{\mathcal{W}_k(x_i)}$. For each word size k, we define the set of all words that exist in x_i but do not exist in y by $\mathcal{R}_k(x_i, \overline{y}) = \mathcal{W}_k(x_i) \cap \overline{\mathcal{W}_k(y)}$. The subset of minimal words as $\mathcal{M}_k(x_i, \overline{y}) = \{\beta \in \mathcal{R}_k(x_i, \overline{y}) : \mathcal{W}_{k-1}(\beta) \cap \mathcal{M}_{k-1}(x_i, \overline{y}) = \emptyset\}$, i.e., a Minimal Absent Word (MAW) of size k cannot contain any MAW of size less than k. In particular, $l\beta r$ is a MAW of sequence x_i, where l and r are single letters from Θ, if $l\beta r$ is not a word of x_i but both $l\beta$ and βr are. We have defined the non-empty set $\mathcal{M}_k(x_i, \overline{y})$ with the smallest k as minimal Relative Absent Words (mRAWs) (Silva *et al.*, 2015).

In this work, we are interested in persistent mRAWs (PmRAWs). Formally, let r be a mRAW of $x_i \in X$ and $P(r, x_i)$ be the predicate "r is a RAW of string x_i". Then, if $\forall_{x_i \in X} P(r, x_i)$, we say that r is persistent in X. This means the full conservation of the identified mRAWs across all the sequences. A particularity of this work is to consider PmRAWs not only at the whole genome level but at a sub-genome level, meaning that a x_i can be considered a subsequence of a whole genome, namely a gene, extending the power of PmRAWs to local observations.

2.2 GC measures

The GC percentage is given by the number of Cytosine (C) and Guanine (G) bases in a string z with length $|z|$ according to

$$\mathcal{GC}(z) = \frac{100}{|z|} \sum_{i=1}^{|z|} \mathcal{N}(z_i || z_i \in \Xi), \qquad (1)$$

where z_i is each symbol of z (assuming causal order), Ξ is a subset alphabet containing the symbols $\{G, C\}$ and $\mathcal{N}$ the program that counts the numbers of symbols in Ξ. Complementary, $\mathcal{AT}(z) = 100 - \mathcal{GC}(z)$. The GC content is obtained with Eq 1 using a sliding window of size 10. Then, the average is computed for all the sequences followed by a low-pass filter with a Hamming window of 20 symbols.

2.3 Alignments and protein structural models

The alignments are built using Bowtie2 (Langmead and Salzberg, 2012), the indexing / sorting with Samtools (Li *et al.*, 2009), and the SNPs map was extracted from IGV (Robinson *et al.*, 2011) according to Supplementary Section 1. The protein structural models were simulated with SWISS-MODEL (Waterhouse *et al.*, 2018) via the ExPASy (Artimo *et al.*, 2012).

3 Results

We improved EAGLE to cope with PmRAWs mapping, GC measures, and visualization. EAGLE is an alignment-free (Zielezinski *et al.*, 2019) tool available at `https://github.com/cobilab/eagle`. The pipeline of the analysis, computer characteristics, and materials used are detailed at Supplementary Section 1, 7, 8, respectively. Although we provide a study on inversions (reverse complementary sequences) in Supplementary Section 3, the results of this paper include models for inversions.

Figure 1-d shows the number of RAWs for lengths between 12 and 17 in SARS-CoV-2 and EBOV (Ebolavirus) outbreaks. As expected, the number of RAWs decreases according to the k-mer size. RAWs for size 11 do not exist in both outbreaks. The number of RAWs is comparatively higher for size 12, 16, and 17 in SARS-CoV-2. On average, there are approximately nine mRAWs in SARS-CoV-2 with size 12 (Supplementary Tables 1 and 2 with averages, variance, and standard deviation for all SARS-CoV-2 and EBOV RAWs). Figure 1-a presents the localization of these RAWs for word lengths of 12, 13, and 14, including the mRAWs (R prefix).

Both R1 and R2 are localized in ORF1a (Figure 1), a non-structural polyprotein involved in the transcription and replication of viral RNAs. It contains the proteinases responsible for the cleavages of the polyprotein. R3 is localized at ORF3, an accessorial protein specialized for environment change inside the infected cell, through the rupture of the membrane, improving the odds of the virus replication. R4 is localized in the Membrane (M), a structural protein that forms part of the outer coat of the virus. It plays a crucial role in virus morphogenesis and assembly via its interactions with other viral proteins. In this case, protein structural prediction was limited (Supplementary Fig. 7). R5 is localized in ORF8, an accessory protein that accumulates several SNPs in SARS-CoV-2 and is dissimilar to other *coronaviruses*. R6 and R7 are localized in the Nucleocapsid protein (N), a structural protein that packages the RNA into a helical ribonucleocapsid (RNP) and is essential during virion assembly through its interactions with the viral genome and membrane protein (M). Also, it enhances the efficiency of subgenomic viral RNA transcription and replication. Protein structural alignments indicate the position of the these RAWs in the simulated structure and the score (Supplementary Fig. 6). R8 and R9 fall within the end of SARS-CoV-2, an ORF that is missing in similar *coronaviruses*.

From these nine mRAWs of size 12, only four are persistent (PmRAW), namely R1 (TGCGCGTCATAT), R2 (GCGCGTCATATT), R4 (TTGCGCGTACGC), and R5 (CGATATCGGTAA). PmRAW R1 and R2 are a super PmRAW because they overlap in eleven symbols, being a PmRAW with size 13 (TGCGCGTCATATT). For a definition of super PmRAW, see Supplementary Section 2. Alignments validate the persistence for the whole 93 SARS-CoV-2 sequences as well as provide how conserved are the PmRAWs flanking regions (Supplementary Fig. 9). All the mRAWs fall in regions without mutations, including SNPs. Additionally, all the mRAWs do not fall in redundant regions that are usually associated with loops, copies, or poly-A tails (Supplementary Fig. 5).

Analysis of mRAWs to other *coronaviruses*, namely SARS-CoV, MERS, HKU, NL63, OC43, 229E, distinguish only one identical mRAW to SARS-CoV-2, R3 (CACAATCGACGG), in SARS-CoV (map with RAWs at Supplementary Fig. 4 and mRAWs described in Supplementary Section 5). The almost nonexistent identity of RAWs between the *coronaviruses* and the fact that R3 is not persistent in SARS-CoV-2 gives an idea of the variability of SARS-CoV-2.

Extended analysis of PmRAWs to local regions (genes/ORFs), according to Figure 1-c and a, reveal the presence of a single PmRAW, RS (CGGCGGGCACGTA), in Spike (S) with size 13. This PmRAW overlaps (only four bases differ in 13) a previous reported SARS-CoV-2 insertion (Andersen *et al.*, 2020). This is a region not present in other *coronaviruses* (CCTCGGCGGGCA), as well as in pangolin, bat, and human genome and

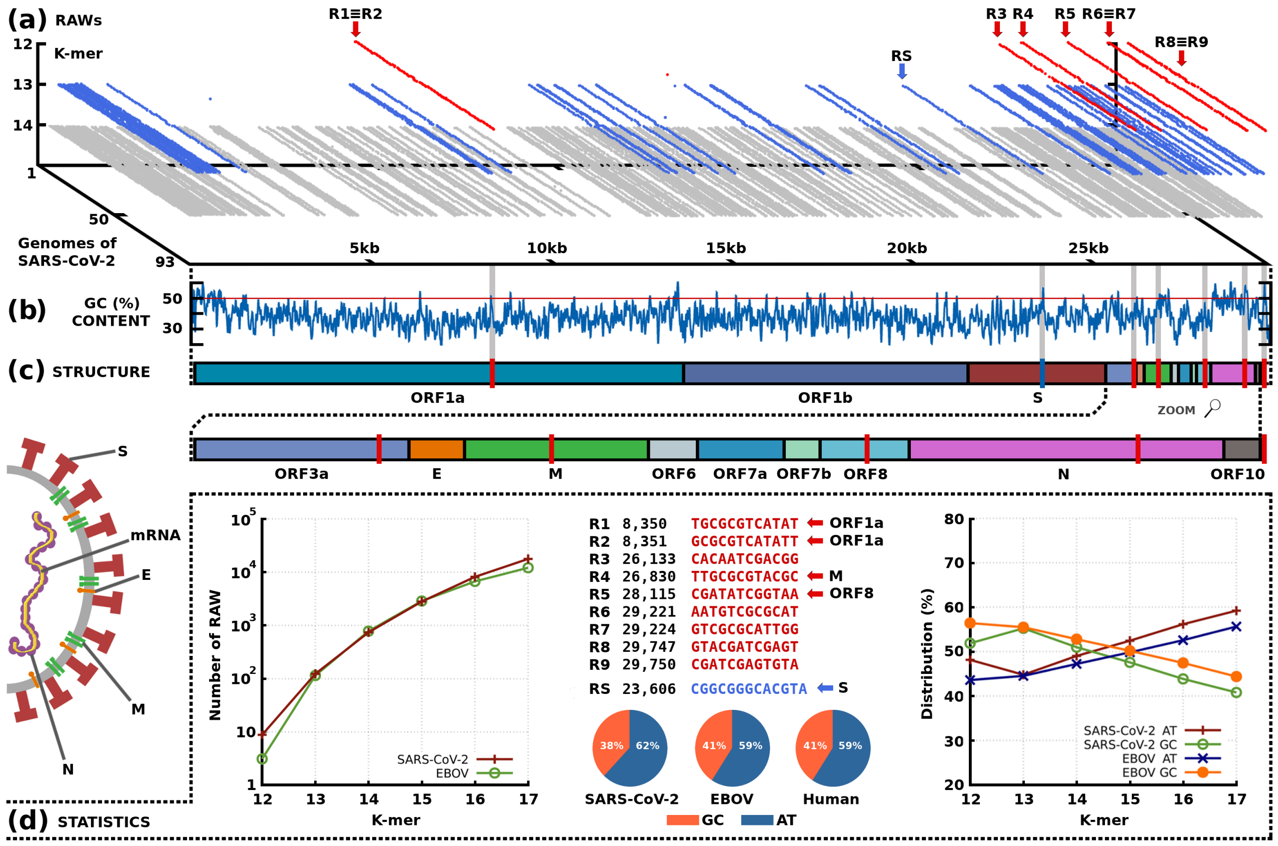

Figure 1. SARS-CoV-2 minimal absent words relative to the complete human genome and transcriptome. (a) identification of RAWs in 93 unaligned genomes from the current outbreak from different countries using EAGLE. RAWs are highlighted in red (k=12, arrows), blue (k=13), and grey (k=14). A vertical projection of this map is in Supplementary Fig. 1. (b) Quantity profiles of guanine/cytosine (GC) for all filtered and averaged whole genomes. (c) Genome structure and extension of the latest regions, where the red stripes stand for the positions of the mRAWs (k=12). (d) Statistics; the left plot depicts the averaged number of RAWs according to different k for SARS-CoV-2 and EBOV; the middle-top identifies RAWs, positions, and the arrows stand for PmRAWs; middle-bottom stands for the proportions of guanine/cytosine (GC) and adenine/thymine (AT) for different genomes (human includes both full genome and transcriptome); the right plot shows the averaged distribution of the GC and AT in different k for two outbreaks. For compatibility purposes, we consider U as T (U of uracil).

transcriptome. Blast analysis of this SARS-CoV-2 signature in the NCBI database, identify *Bradymonadales bacterium* as the highest match. Using protein structural analysis, we show that this motif is exposed at the protein surface (Supplementary Fig. 8), yet its RNA is persistent.

SARS-CoV-2 genomes are characterized by a low GC percentage ($\approx 38\%$), lower than EBOV ($\approx 41\%$) and human ($\approx 42\%$) (Figure 1-d). In Figure 1-b, the GC content reflects an average profile with minimal regions above 50%. We found that mRAWs have higher GC% than the average genome, while RAWs with higher size (k-mer > 14) progressively approximate to the genome $GC\%$. Similar results are present in EBOV and other human *coronaviruses* (Figure 1-d and Supplementary Fig. 10).

4 Discussion

In this article, we revealed the existence of short DNA/RNA sequences, with sizes comprehended between 12 and 17, that are present in 93 SARS-CoV-2 genomes and absent from the human genome and transcriptome, including possible inversions.

The analysis exposed a new pattern related to the distribution of the RAWs, characterized by an increasing GC content, according to the decrease of RAWs length, yielding the same behavior in additional viral genomes, namely human *coronavirus* and 159 Ebolavirus. Given the high energy-richness (Zhang *et al.*, 2004) of the mRAWs and the contrary average GC distributions of both viral and host genomes, these regions are possibly uncovering the presence of higher dimensional structures not featured in the host. This pattern opens a future research line and strives for confirmation of identical nature in single and double-strand DNA viruses.

We have defined a particular RAWs subset, namely the shortest sequences that are persistent across all the genomes (PmRAWs). We identified such sequences, in particular, four sequences of size twelve and one with size thirteen. The identified PmRAWs describe minimal signatures of the SARS-CoV-2 genome that provides distinguishability between human *coronavirus* species and useful identifiability characteristics.

Efficient diagnostic methods are crucial in isolation for this pandemic. Despite the current race in antibodies research and their federal approval, which will allow testing the immunity answer of a subject regarding SARS-CoV-2, the standard diagnostic method is real-time RT-PCR (Reverse Transcriptase quantitative Polymerase Chain Reaction) applied to subregions of SARS-CoV-2. However, the sensitivity of RT-PCR is limited, given a considerable number of possible false negatives (Nalla *et al.*, 2020). As a consequence, uncertainty increases between possible reinfections and interleaved outcomes given by the method.

The evolution of SARS-CoV-2 is supplied by the replication of RNA sequences with mutations, mostly SNPs, harming the RT-PCR method's sensitivity. The identification of PmRAWs in SARS-CoV-2 allows enhancing the RT-PCR methodologies through the usage of conserved regions between multiple genomes of SARS-CoV-2. The reported PmRAWs are absent from other human coronaviruses and, hence, diminish the probability of cross-reaction. Moreover, the identification of persistent sequences that are absent from the human genome and transcriptome permits the isolation of the SARS-CoV-2 signal from possible human material that abounds in the cell in higher magnitude orders of quantity. Therefore, PmRAW sequences can be used in diagnosis to design primers that identify SARS-CoV-2 infections or distinguish between *coronavirus* species.

With a high number of COVID-19 patients suffering from severe disease and hospitals being overwhelmed, treatments are urgently needed. Since creating new compounds may take years to develop and test, besides general strategies to improve the immune system, the majority of the candidate therapeutics are based on the search for the most effective drugs developed in previous contexts or successfully in-vitro. Some examples are plasma transfusions, Chloroquine, hydroxychloroquine, Remdesivir, and Ritonavir/lopinavir (Li *et al.*, 2020). Some of these are already at phase of dosing optimization and randomized controlled trials.

On the other hand, novel therapeutics are also being investigated and tested. Specifically, the SARS-Cov-2 Spike (S) glycoprotein is being used as a target for vaccines, namely because Spike binds into the angiotensin-converting enzyme (ACE) 2 human receptor for a further injection of the viral mRNA into the human cell (Wrapp *et al.*, 2020). The identified PmRAW (RS) may be a key for a potential target therapeutic, given its singularity, persistency, and exposer at the Spike protein surface.

The PmRAWs with size 12 are also potential targets. The overlap of R1 with R2 permits to decrease the likelihood of interference regarding possible human or viral mutations. R5 seems to be more limited, given the high SNPs intensity flanking the RAW. Contrarily, R4 appears within the sequence of the viral Membrane (M) protein. The M outside exposer of the virus (Figure 1-c left) constitutes an appealing target.

The identified signatures of SARS-CoV-2 in this article, provide useful means for specific characterization and the potential for the emergence of new therapeutic applications, that require mandatory *in-vitro* and *in-vivo* implementation and testing for accurate verification.

Funding

D.P. is funded by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the Scientific Employment Stimulus - Institutional Call - CI-CTTI-94-ARH/2019. J.M.S. acknowledges the FCT grant SFRH/BD/141851/2018. We also acknowledge national Funds through the FCT in the context of the project UIDB/00127/2020.

References

Andersen, K. G. *et al.* (2020). The proximal origin of SARS-CoV-2. *Nature Medicine*, pages 1–3.

Artimo, P. *et al.* (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic acids research*, **40**(W1), W597–W603.

Barton, C. *et al.* (2014). Linear-time computation of minimal absent words using suffix array. *BMC bioinformatics*, **15**(1), 388.

Béal, M.-P. *et al.* (1996). Minimal forbidden words and symbolic dynamics. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 555–566. Springer.

Chairungsee, S. and Crochemore, M. (2012). Using minimal absent words to build phylogeny. *Theoretical Computer Science*, **450**, 109–116.

Chen, N. *et al.* (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, **395**(10223), 507–513.

Crochemore, M. *et al.* (1998). Automata and Forbidden Words. *Inf. Process. Lett.*, **67**(3), 111–117.

Crochemore, M. *et al.* (2020). Absent words in a sliding window with applications. *Information and Computation*, **270**, 104461.

Drosten, C. *et al.* (2003). Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *New England journal of medicine*, **348**(20), 1967–1976.

Falda, M. *et al.* (2014). keeSeek: searching distant non-existing words in genomes for PCR-based applications. *Bioinformatics*, **30**(18).

Garcia, S. P. *et al.* (2011). Minimal absent words in prokaryotic and eukaryotic genomes. *PLoS One*, **6**(1).

Gurwitz, D. (2020). Angiotensin receptor blockers as tentative SARS-CoV-2 therapeutics. *Drug development research*.

Héliou, A. *et al.* (2017). emMAW: computing minimal absent words in external memory. *Bioinformatics*, **33**(17), 2746–2749.

Herold, J. *et al.* (2008). Efficient computation of absent words in genomic sequences. *BMC bioinformatics*, **9**(1), 167.

Huang, C. *et al.* (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, **395**(10223), 497–506.

Lai, C.-C. *et al.* (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *International journal of antimicrobial agents*.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, **9**(4), 357.

Li, H. *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, H. *et al.* (2020). Updated approaches against SARS-CoV-2. *Antimicrobial agents and chemotherapy*, **64**(6).

Nalla, A. K. *et al.* (2020). Comparative performance of SARS-CoV-2 detection assays using seven different primer-probe sets and one assay kit. *Journal of clinical microbiology*, **58**(6).

Nguyen, T. M. *et al.* (2020). Virus against virus: a potential treatment for 2019-nCov (SARS-CoV-2) and other RNA viruses.

Pinho, A. J. *et al.* (2009). On finding minimal absent words. *BMC bioinformatics*, **10**(1), 137.

Robinson, J. T. *et al.* (2011). Integrative genomics viewer. *Nature biotechnology*, **29**(1), 24–26.

Santoni, D. and Vergni, D. (2020). In the search of potential epitopes for Wuhan seafood market pneumonia virus using high order nullomers. *Journal of Immunological Methods*, page 112787.

Silva, R. M. *et al.* (2015). Three minimal sequences found in Ebola virus genomes and absent from human DNA. *Bioinformatics*, **31**(15).

Vergni, D. and Santoni, D. (2016). Nullomers and high order nullomers in genomic sequences. *PloS one*, **11**(12).

Wang, C. *et al.* (2020). A novel coronavirus outbreak of global health concern. *The Lancet*, **395**(10223), 470–473.

Waterhouse, A. *et al.* (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, **46**(W1), W296–W303.

Wrapp, D. *et al.* (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, **367**(6483), 1260–1263.

Wu, F. *et al.* (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, **579**(7798), 265–269.

Wu, Z.-D. *et al.* (2010). Efficient computation of shortest absent words in a genomic sequence. *Information Processing Letters*, **110**(14-15).

Zaki, A. M. *et al.* (2012). Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *New England Journal of Medicine*, **367**(19), 1814–1820.

Zhang, H. *et al.* (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine*, pages 1–5.

Zhang, L. *et al.* (2004). GC/AT-content spikes as genomic punctuation marks. *Proceedings of the National Academy of Sciences*, **101**(48), 16855–16860.

Zhou, P. *et al.* (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**(7798), 270–273.

Zielezinski, A. *et al.* (2019). Benchmarking of alignment-free sequence comparison methods. *Genome biology*, **20**(1), 144.