

Severe acute respiratory syndrome coronavirus 2 may exploit human transcription factors involved in retinoic acid and interferon-mediated response: a hypothesis supported by an *in silico* analysis

I. di Bari¹, R. Franzin¹, A. Picerno¹, A. Stasi¹, M. T. Cimmarusti¹, M. Di Chiano¹, C. Curci^{1,2}, P. Pontrelli¹, M. Chironna⁴, G. Castellano⁵, A. Gallone², C. Sabbà³, L. Gesualdo¹ and F. Sallustio³

1) Department of Emergency and Organ Transplantation, 2) Department of Basic Medical Sciences, Neuroscience and Sense Organs, 3) Department of Interdisciplinary Medicine, University of Bari 'Aldo Moro', 4) Department of Biomedical Sciences and Human Oncology– Hygiene Section, University of Bari, Bari and 5) Department of Medical and Surgical Science, University of Foggia, Foggia, Italy

Abstract

The pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing coronavirus disease 2019 (COVID-19), resulting in acute respiratory disease, is a worldwide emergency. Because recently it has been found that SARS-CoV is dependent on host transcription factors (TF) to express the viral genes, efforts are required to understand the molecular interplay between virus and host response. By bioinformatic analysis, we investigated human TF that can bind the SARS-CoV-2 sequence and can be involved in viral transcription. In particular, we analysed the key role of TF involved in interferon (IFN) response. We found that several TF could be induced by the IFN antiviral response, specifically some induced by IFN-stimulated gene factor 3 (ISGF3) and by unphosphorylated ISGF3, which were found to promote the transcription of several viral open reading frame. Moreover, we found 22 TF binding sites present only in the sequence of virus infecting humans but not bat coronavirus RaTG13. The 22 TF are involved in IFN, retinoic acid signalling and regulation of transcription by RNA polymerase II, thus facilitating its own replication cycle. This mechanism, by competition, may steal the human TF involved in these processes, explaining SARS-CoV-2's disruption of IFN-I signalling in host cells and the mechanism of the SARS retinoic acid depletion syndrome leading to the cytokine storm. We identified three TF binding sites present exclusively in the Brazilian SARS-CoV-2 P.1 variant that may explain the higher severity of the respiratory syndrome. These data shed light on SARS-CoV-2 dependence from the host transcription machinery associated with IFN response and strengthen our knowledge of the virus's transcription and replicative activity, thus paving the way for new targets for drug design and therapeutic approaches.

© 2021 The Authors. Published by Elsevier Ltd.

Keywords: Interferon regulatory factors, ISGF3, retinoic acid, SARS-CoV-2

Original Submission: 5 January 2021; **Revised Submission:** 22 February 2021; **Accepted:** 24 February 2021

Article published online: 27 February 2021

Corresponding author: F. Sallustio, Department of Interdisciplinary Medicine, University of Bari 'Aldo Moro', Piazza G. Cesare, 11, Bari, 70124, Italy.

E-mail: fabio.sallustio@uniba.it

The last two authors contributed equally to this article, and both should be considered senior author.

Introduction

Coronaviruses (CoV) are a large family of respiratory viruses that can cause mild to moderate diseases, from the common cold to respiratory syndromes such as Middle East respiratory syndrome (MERS) and severe acute respiratory syndrome (SARS). CoVs are common in many animal species (such as camels and bats), but in some cases, although rarely, they can evolve and infect humans, then spread throughout the population [1,2].

In 2019, a new species of CoV never before identified in humans, severe acute respiratory syndrome coronavirus 2

(SARS-CoV-2), was recognized. It is an RNA virus coated with a capsid and a pericapsid crossed by glycoprotein structures that give it the typical 'crown' appearance; it is 100 to 160 nm in diameter [3]. In humans, SARS-CoV-2 is capable of causing CoV disease 2019 (COVID-19), a severe acute respiratory syndrome that leads to a high mortality rate and several complications, including acute kidney injury in about 25% of patients [4].

The scientific community is currently trying to identify the way the virus is transmitted to humans [5]. The SARS-CoV-2 genome is made up of a single strand of large-size positive RNA (+ssRNA), about 30 kb in size, which gives rise to seven viral proteins and is associated with protein N, which increases its stability (Fig. 1(A)).

It has long been established that RNA viruses frequently subvert cellular factors for replication, and transcription of viral RNAs and viral promoters/enhancers should be activated by the same signalling events as innate immune genes. Therefore, cellular factors are active in the replication and transcription of viral RNAs [6]. CoV infection affects both host cell transcription and translation, and both viral and cellular proteins are required for replication and transcription [7,8]. Recently it has also been experimentally demonstrated how extensively the TF host is involved in MERS-CoV replication/transcription activity [9].

We investigated human TF that can bind the SARS-CoV-2 sequence and found that several TF can be induced by the interferon (IFN) antiviral response. This led to hypothesize that the virus may use this pathway to modulate the cellular RNA polymerase.

Materials and methods

Selection of CoV sequences for comparison to COVID-19

All full-length sequences from SARS-CoV Italian patients were retrieved from the ViPR (<https://www.viprbrc.org/brc/home.spg?decorator=corona>) and GISAID (<https://www.gisaid.org/>) databases on 30 March 2020. Sequences considered had the following IDs: EPI_ISL_412974; EPI_ISL_413489; EPI_ISL_417419; EPI_ISL_417418; EPI_ISL_417423; EPI_ISL_417421; EPI_ISL_412973; and EPI_ISL_417447. British B.1.1.7 (EPI_ISL_601443), Marseille (EPI_ISL_569131), Brazilian P.1 (EPI_ISL_833137) and African 501Y.V2 (EPI_ISL_712073) sequences were used for the comparisons. Sequences were aligned using the MUSCLE algorithm in ViPR. The consensus sequence of Italian virus was determined from the final alignments using the sequence variation analysis tool in ViPR. The *Coronaviridae* ViPR database was used to identify similar sequences using the BLAST (Basic Local Alignment

Search Tool; <https://blast.ncbi.nlm.nih.gov/Blast.cgi>) algorithm to compute the pairwise identity between Italian consensus sequence and its comparison target.

Identification of TF binding SARS-CoV-2 sequence

To identify tumor-factor binding to the SARS-CoV-2 sequence, we used CiiDER scanning [10], which utilizes an implementation of the MATCH algorithm [11] to predict potential tumor-factor binding sites (TFBS) in regions of interest. A deficit value of 0.15, which is the difference between the MATCH score of a TFBS and the maximum possible score, which is 1, was used. For the scanning, position frequency matrices (PFMs) from the curated open-access JASPAR database (<http://jaspar.genereg.net/>) [12,13] were used. In order to predict whether a sequence contains a site for a TF, CiiDER compared the virus sequence to the PFM and generated a score of similarity. If the sequence matches the PFM perfectly, then the deficit value is 0. To search for TFBS, sequences are split into overlapping five-base regions, which are compared with the core PFM. If the similarity score between a five-base sequence and the core PFM meets a defined threshold, then the sequence window is increased to the full length of the TFBS and the similarity score with the full PFM is calculated.

For the scans, we used the positive strand for open reading frames (ORF) 1a and b, and the negative strand for the remaining genes and ORF. An enrichment analysis, comparing the distribution of TFBS predicted in a set of regulatory regions to the distribution in a set of background sequences, has been utilized to more accurately identify TFBS that are statistically over- or underrepresented [14]. Over- and underrepresented TF were determined by comparing the numbers of sequences with predicted TFBS to the number of those without by Fisher exact test. As background sequences, we used binding sequences of TF induced by ISGF3 and by unphosphorylated ISGF3 (U-ISGF3) [15].

Results

Human SARS-CoV-2 has 96% similarity with bat SARS-like CoV RaTG13

The sequences of SARS-CoV-2 from eight different Italian patients were multiply aligned and compared to identify a consensus sequence for the Italian viral species. Then the consensus sequence was compared to all sequences included in the *Coronaviridae* ViPR database to identify similar sequences. The databases included 198 species, 6938 genomes and 6642 strains. We found that our consensus sequence had a similarity of 88% with a bat SARS-like CoV isolated in China,

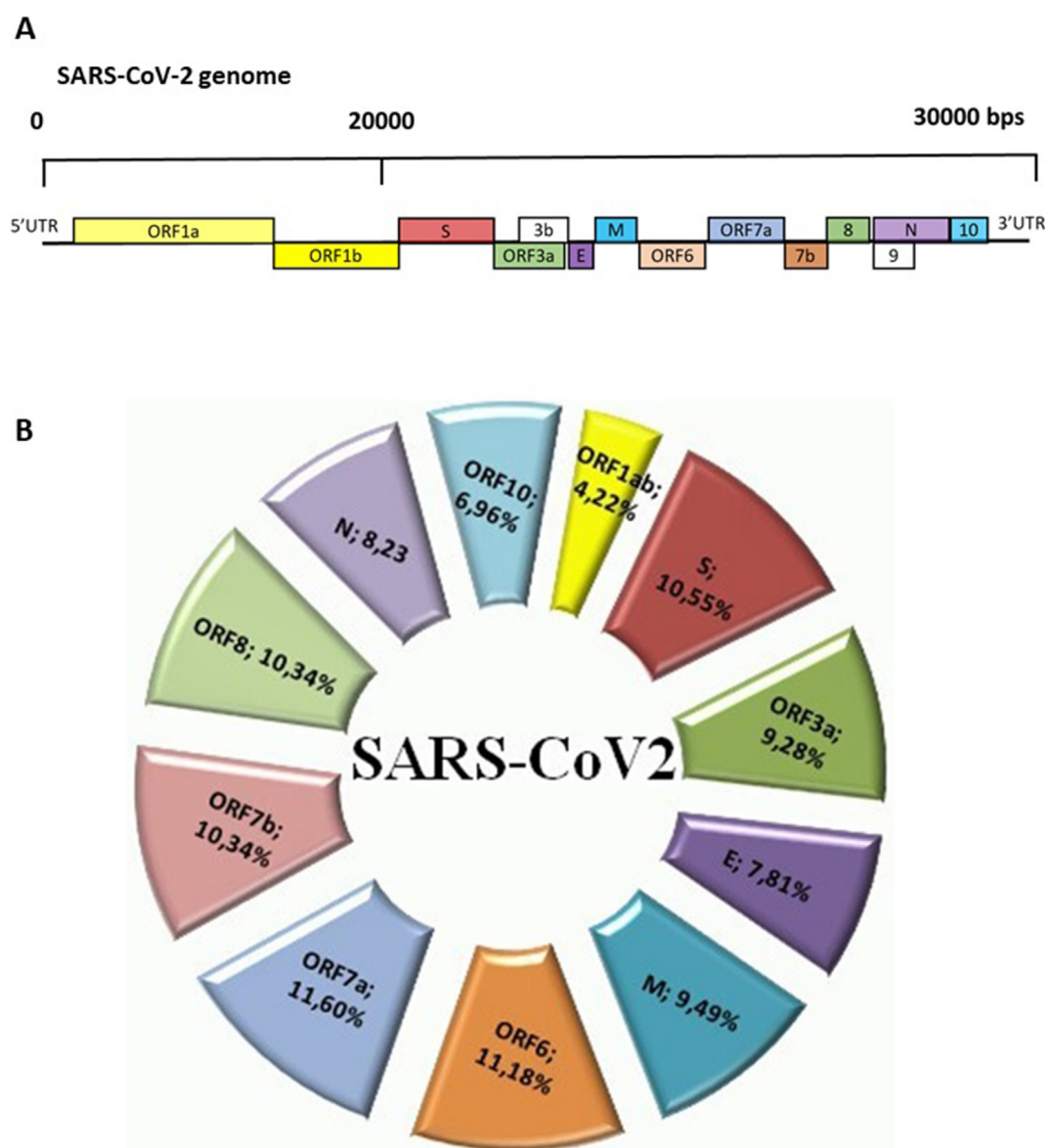


FIG. 1. Schematic overview of genome, genes and proteins of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). (A) SARS-CoV-2 genome comprises positive-sense, single-stranded RNA (ssRNA) genome 27 to 32 kb in size. The 5' terminus (translated from first ORF1a and ORF1b) encodes two large polyproteins, pp1a and pp1ab, which are proteolytically cleaved into 16 nonstructural proteins (NSPs), including papain-like protease (PLpro), 3C-like protease (3CLpro) and RNA-dependent RNA polymerase (RdRp). An additional 9 to 12 open reading frames (ORFs) are encoded through transcription of nested set of subgenomic RNAs. The 3' terminus encodes structural proteins, including envelope glycoproteins spike (S), envelope (E), membrane (M) and nucleocapsid (N). (B) Percentage distribution of transcription factor (TF) binding sites in different genomic regions of SARS-CoV-2, as follows: ORF1ab, open reading frame 1ab; S, protein S; ORF3a, open reading frame 3a; E, protein E; M, protein M; ORF6, open reading frame 6; ORF7a, open reading frame 7a; ORF7b, open reading frame 7b; ORF8, open reading frame 8; N, protein N; ORF10, open reading frame 10.

[gb]AVP78030 (score = 1.995e+04 bits). Moreover, we compared the consensus sequence with a closely related bat RaTG13 (MN996532.2) isolated from *Rhinolophus affinis* (horseshoe bat), and we found a similarity of 96.14% (score 4.875e+04).

Human TF is potentially involved in SARS-CoV-2 transcription

By using the Italian consensus sequence, we assessed the possible human TF able to bind the viral RNA. We considered the JASPAR vertebrate database. This model represents the

majority of human TF. A total of 659 TF belonging to the JASPAR database and binding to sequences of SARS-CoV-2 was found (Supplementary Table S1). Then we selected TF that bind the viral sequence near the starting site of each of the viral gene or ORF considering a range of ± 1000 bases from the starting sites (Supplementary Table S2). We found that the binding sites for TF were equally distributed between the promoter regions of each transcribed sequence except for ORF1ab and ORF10, in which TFBS were underrepresented (Fig. 1(B)). Only 4.2% of the TF can bind ORF1ab and 6.7% can bind ORF10.

TF is induced by IFN

We performed an enrichment analysis to identify TF induced by IFN and whose binding sites were overrepresented in the Italian SARS-CoV-2 sequence. In particular, we investigated TF involved in the activation of IFN-stimulated gene factor 3 (ISGF3), which drives the expression of more than 100 IFN- β -stimulated genes [16]. We subdivided genes and TF induced by ISGF3 and those induced by U-ISGF3, which maintains the expression of a subset of the initially induced IFN-stimulated genes (ISGs), whose protein products lead to extended resistance to virus infection and DNA damage. We found that SARS-CoV-2 contains 18 binding sites for 11 ISGF3-induced TF located near the transcription starting sites of virus genes

(Table 1 and Fig. 2). Most of ISGF3-induced TF bound promoters of S protein and ORF3a, whereas no TF bound ORF1ab (Fig. 2). Moreover, the same TF, Six1, could bind the promoter regions of genes coding for structural proteins E and M and of ORF6, -7, -8 and -9. However, 36 different binding sites for 19 U-ISGF3-induced TF were present (Table 2). Among these, seven different U-ISGF3-induced TF bound ORF3a, four TF bound the promoter of the gene coding for the S protein and three TF bound ORF1ab. SOX21 and HOXA9 could bind three different promoter regions of ORF3 and genes E and M, whereas FOXP1 and Hmx1 could bind five different promoter regions of ORF6, -7, -8 and -9 and N gene. Among U-ISGF3-induced TF is included again SIX1, which binds promoter regions of four ORFs and two genes of the virus (Fig. 3).

Polymorphisms affecting TFBS appear in English, French, African and Brazilian SARS-CoV-2 variants

We performed the host TFBS analysis on English B.1.1.7 (EPI_ISL_601443), Marseille (EPI_ISL_569131), South African 501Y.V2 (EPI_ISL_712073) and Brazilian (P.1) SARS-CoV-2 variants. We did not find polymorphisms affecting TFBS in English, Marseille or South African virus variants, whereas interestingly, we found two mutations affecting TFBS within the Brazilian variant P.1. The mutation 26150 T/C causes the

TABLE 1. SARS-CoV-2 binding sites for ISGF3-induced TF

Gene	Start	TF name	TF ID	Start position	End position	Core match score	Matrix match score	Sequence
orf1ab	266	POU4F2	MA0683.1	125	140	1	0.858	CAGTATAATTAATAAC
orf1ab	266	RFX2	MA0600.2	659	674	0.89	0.855	GGTGGCCATAGTTACG
orf1ab	266	XBPI	MA0844.1	385	398		0.95	AGAGGCACGTCAC
S	21563	HOXD9	MA0913.2	21578	21587		1	GCAATAAAC
S	21563	ATOH1 (var. 2)	MA1467.1	22336	22345		0.957	ACCAGCTGTC
S	21563	HOXD13	MA0909.2	21578	21588		0.993	GGCAATAAAC
S	21563	Stat5a	MA1624.1	21178	21189		0.999	ATTCCAAGAATG
S	21563	CDX4	MA1473.1	21578	21588		1	GGCAATAAAC
ORF3a	25393	NEUROG2	MA0669.1	24577	24586		0.957	CACATATGTC
ORF3a	25393	Six3	MA0631.1	24491	24507		0.904	GAAAGGATATCATTTAA
ORF3a	25393	ZSCAN4	MA1155.1	24647	24661		0.888	TACACACTCTGACAT
ORF3a	25393	NEUROG1	MA0623.2	24577	24586		0.971	CACATATGTC
ORF3a	25393	Alx1	MA0854.1	24591	24607		0.98	AGCTCTAATTAATTGTT
ORF3a	25393	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF3a	25393	HOXA9	MA0594.2	25848	25857		0.99	GTCGTAACAA
E	26245	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
E	26245	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
E	26245	HOXA9	MA0594.2	25848	25857		0.99	GTCGTAACAA
M	26523	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
M	26523	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
M	26523	HOXA9	MA0594.2	25848	25857		0.99	GTCGTAACAA
ORF6	27202	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF6	27202	FOXP1	MA0852.2	28136	28149		0.993	AAGGTAACAGGAA
ORF6	27202	Hmx1	MA0896.1	28151	28167		0.955	CCTGGCAATTAATTGTA
ORF7a	27394	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7a	27394	FOXP1	MA0852.2	28136	28149		0.993	AAGGTAACAGGAA
ORF7a	27394	Hmx1	MA0896.1	28151	28167		0.955	CCTGGCAATTAATTGTA
ORF7b	27756	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7b	27756	FOXP1	MA0852.2	28136	28149		0.993	AAGGTAACAGGAA
ORF7b	27756	Hmx1	MA0896.1	28151	28167		0.955	CCTGGCAATTAATTGTA
ORF8	27894	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF8	27894	FOXP1	MA0852.2	28136	28149		0.993	AAGGTAACAGGAA
ORF8	27894	Hmx1	MA0896.1	28151	28167		0.955	CCTGGCAATTAATTGTA
N	28274	FOXP1	MA0852.2	28136	28149		0.993	AAGGTAACAGGAA
N	28274	Hmx1	MA0896.1	28151	28167		0.955	CCTGGCAATTAATTGTA
ORF10	29558	MAFA	MA1521.1	29327	29341		0.923	ATGCTTATTCAGCAA

Abbreviations: ISGF3, interferon-stimulated gene factor 3; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; TF, transcription factor.

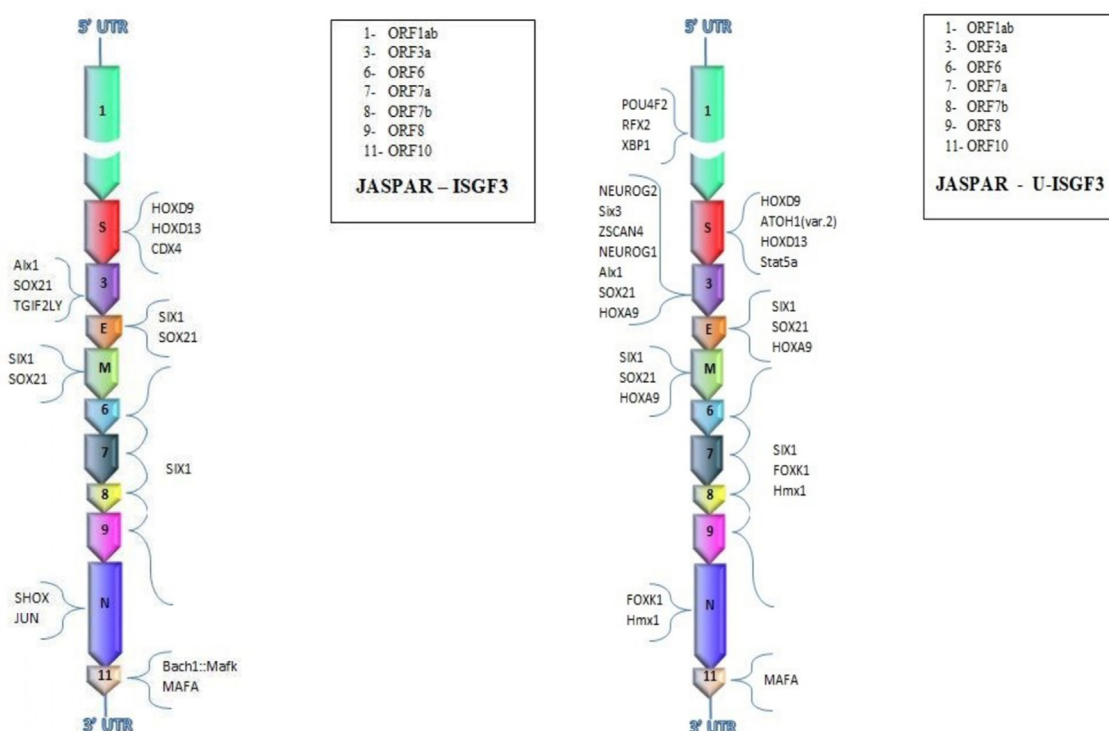


FIG. 2. Interferon-stimulated gene factor 3 (ISGF3)-induced and unphosphorylated ISGF3 (U-ISGF3)-induced transcription factors (TF) with binding sites present in different genomic regions of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), as follows: (1) ORF1ab, open reading frame 1ab; S, protein S. (3) ORF3a, open reading frame 3a; E, protein E; M, protein M. (6) ORF6, open reading frame 6. (7) ORF7a, open reading frame 7a. (8) ORF7b, open reading frame 7b. (9) ORF8, open reading frame 8; N, protein N. (11) ORF10, open reading frame 10.

formation of the binding site for the SPEDEF TF in the P.I sequence (Fig. 3(A)). Moreover, the enrichment analysis to identify TF induced by IFN and whose binding sites were overrepresented in the Brazilian (P.I) SARS-CoV-2 sequence showed that the mutation 6320 A/G leads to enrichment for the binding sites of GRHL1 and TFCEP2 TF (Fig. 3(B) and (C)). All remaining TFBS were shared by all four SARS-CoV-2 variants.

Binding sites of TF involved in retinoic acid metabolism and in transcription regulation by RNA polymerase II are present exclusively in SARS-CoV-2 genome isolated from humans

We then compared TFBS present in the SARS-CoV-2 sequence isolated from Italian patients (Supplementary Table S1) and those present in the sequence of bat CoV RaTG13 (Supplementary Table S3). Interestingly, we found that 22 TFBS were present exclusively in the genomic region of SARS-CoV-2 isolated from humans and not in the virus isolated from bats (Table 3); among these, six TFBS were near the starting site of each of the viral gene or ORF considering a range of ± 1000 bases from the starting sites (Supplementary Table S4).

Moreover, we performed a network analysis and found that the 22 TF formed a significant network (8.27e-13, Fig. 4) showing the involvement of TF in the retinoic acid signalling through the direct binding of the retinoic acid receptor RXR- α gene by NR1H3, KLF5, THRB and PAX6 TF (Fig. 4). In addition, the functional enrichment analysis showed that most significant biological processes in which these genes are involved were constituted by the regulation of transcription by RNA polymerase II (false discovery rate = 9.88e-05).

Discussion

The CoV RNA synthesis is performed by a multienzymatic replicase complex together with cellular factors; the process requires the specific recognition of RNA *cis*-acting signals located in the virus genome. Cellular proteins are involved in CoV RNA synthesis together with the p100 transcriptional coactivator protein [8,17]. A strong interaction between host cell and virus replication and transcription processes exists in viruses of the *Coronaviridae* family [18]. In the absence of a specific mechanism for the control of the cell and its replicative

TABLE 2. SARS-CoV-2 binding sites for unphosphorylated ISGF3-induced TF

Gene	Start	TF name	TF ID	Start position	End position	Core match score	Matrix match score	Sequence
S	21563	HOXD9	MA0913.2	21578	21587			GCAATAAAAC
S	21563	HOXD13	MA0909.2	21578	21588		0.993	GGCAATAAAAC
S	21563	CDX4	MA1473.1	21578	21588			GGCAATAAAAC
ORF3a	25393	Alx1	MA0854.1	24591	24607		0.98	AGCTCTAATTAATTGTT
ORF3a	25393	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF3a	25393	TGIF2LY	MA1572.1	24799	24810		0.896	TGACAAATGGCA
E	26245	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
E	26245	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
M	26523	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
M	26523	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF6	27202	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7a	27394	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7b	27756	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF8	27894	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
N	28274	SHOX	MA0630.1	28490	28497			TTAATTGG
N	28274	JUN	MA0488.1	27290	27302		0.965	TTTATGATGTAAT
ORF10	29558	Bach1::Mafk	MA0591.1	29511	29525		0.923	GAGTTGAGTCAGCAC
ORF10	29558	MAFA	MA1521.1	29327	29341		0.923	ATGCTTATTCAGCAA
S	21563	HOXD9	MA0913.2	21578	21587			GCAATAAAAC
S	21563	HOXD13	MA0909.2	21578	21588		0.993	GGCAATAAAAC
S	21563	CDX4	MA1473.1	21578	21588			GGCAATAAAAC
ORF3a	25393	Alx1	MA0854.1	24591	24607		0.98	AGCTCTAATTAATTGTT
ORF3a	25393	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF3a	25393	TGIF2LY	MA1572.1	24799	24810		0.896	TGACAAATGGCA
E	26245	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
E	26245	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
M	26523	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
M	26523	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF6	27202	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7a	27394	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7b	27756	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF8	27894	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
N	28274	SHOX	MA0630.1	28490	28497			TTAATTGG
N	28274	JUN	MA0488.1	27290	27302		0.965	TTTATGATGTAAT
ORF10	29558	Bach1::Mafk	MA0591.1	29511	29525		0.923	GAGTTGAGTCAGCAC
ORF10	29558	MAFA	MA1521.1	29327	29341		0.923	ATGCTTATTCAGCAA
S	21563	HOXD9	MA0913.2	21578	21587			GCAATAAAAC
S	21563	HOXD13	MA0909.2	21578	21588		0.993	GGCAATAAAAC
S	21563	CDX4	MA1473.1	21578	21588			GGCAATAAAAC
ORF3a	25393	Alx1	MA0854.1	24591	24607		0.98	AGCTCTAATTAATTGTT
ORF3a	25393	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF3a	25393	TGIF2LY	MA1572.1	24799	24810		0.896	TGACAAATGGCA
E	26245	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
E	26245	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
M	26523	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
M	26523	SOX21	MA0866.1	26089	26103		0.91	AACAATTTTATTGTA
ORF6	27202	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7a	27394	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF7b	27756	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
ORF8	27894	SIX1	MA1118.1	27220	27230		0.995	GTAACCTGAAA
N	28274	SHOX	MA0630.1	28490	28497			TTAATTGG
N	28274	JUN	MA0488.1	27290	27302		0.965	TTTATGATGTAAT
ORF10	29558	Bach1::Mafk	MA0591.1	29511	29525		0.923	GAGTTGAGTCAGCAC
ORF10	29558	MAFA	MA1521.1	29327	29341		0.923	ATGCTTATTCAGCAA

Abbreviations: ISGF3, interferon-stimulated gene factor 3; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; TF, transcription factor.

machinery for the synthesis of viral products, CoV infection affects both host cell transcription and translation, and both viral and cellular proteins are required for replication and transcription [7].

Recently it has been experimentally demonstrated how extensively the host is involved in the MERS-CoV replication/transcription activity. Indeed, more than 500 host proteins constituting the replication/transcription complex microenvironment have been identified [9]. Among these biochemically validated cellular factors, eight TF which we identified in our analysis were present (STAT1, STAT3, TFEB, NFIX, NFIC, Stat5a, NR3C1, and Zfx).

Viral immediate early control elements might functionally mimic innate immune enhancers, gaining advantage from TF

activated by immune signalling to induce viral immediate early gene expression [19]. Here we found that SARS-CoV-2 genome is enriched in binding sites for TF activated by IFN- β . IFNs such as IFN- α , IFN- β and IFN- γ are important antiviral cytokines released during infection. In the case of these specific viruses, the IFN response process also facilitates virus replication and gene expression, as in HIV (+ssRNA) [20,21].

The CoV nonstructural proteins, in collaboration with recruited host cell proteins, constitute membrane-associated replication and transcription complexes. SARS-CoV nsp1 was the first CoV nsp1 demonstrated to block the expression of reporter gene under the control of a IFN- β promoter [22]. From our analysis, a key role of some TF, such as the family of IFN regulatory factors (IRFs), has emerged.

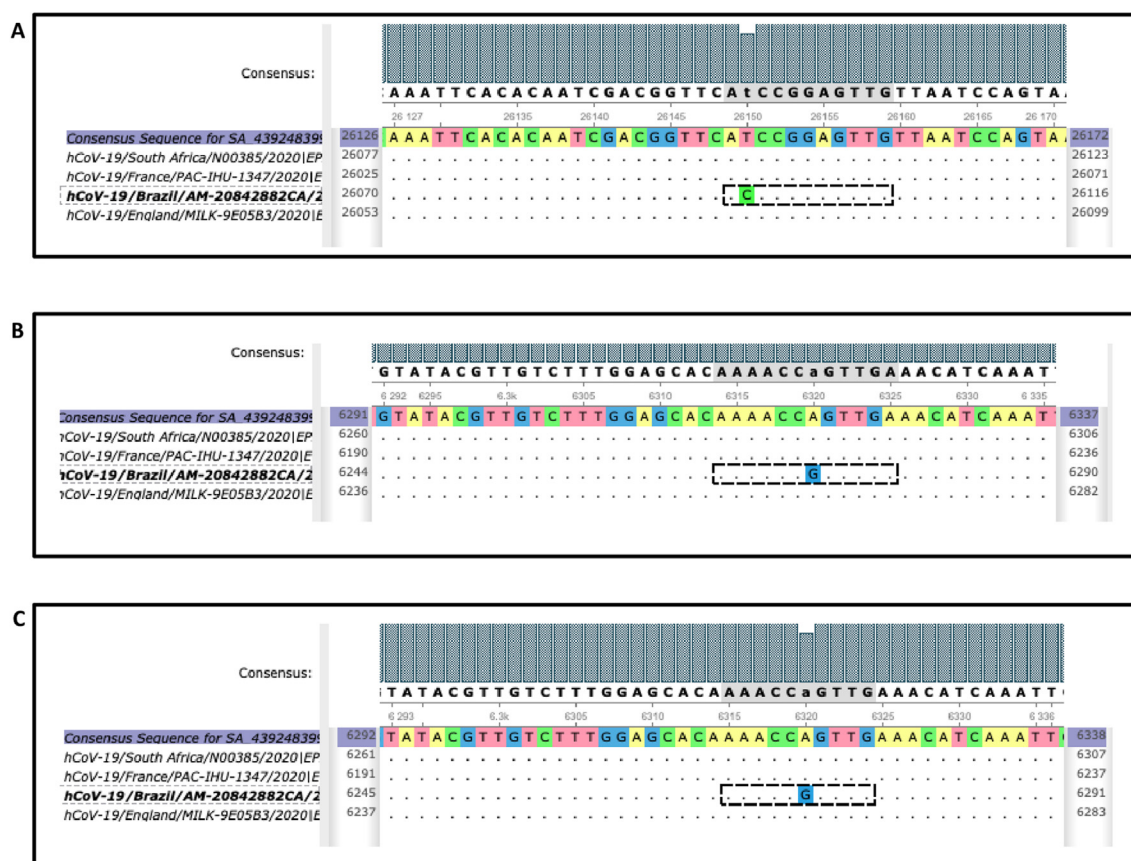


FIG. 3. Polymorphisms affecting transcription factor (TF) binding sites in P.I Brazilian severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variant. (A) Mutation 26150 T/C causes formation of binding site for SPEDEF TF in P.I sequence. (B, C) Mutation 6320 A/G leads to enrichment for binding sites of GRHL1 and TFCP2 TF.

In particular, we found two sets of TF. One is induced by the TF ISGF3 (IRF9 and tyrosine-phosphorylated STAT1/2), which drives the first rapid cellular response to the viral infection; the other is induced by the related factor U-ISGF3, which drives the second prolonged response. U-ISGF3 is composed by IFN- β -induced IRF9 and STAT1/2 without tyrosine phosphorylation. The U-ISGF3-induced antiviral genes that show prolonged expression are driven by distinct IFN-stimulated response elements. Interestingly, our data are also supported by the recent discovery that the SARS-CoV-2 receptor angiotensin-converting enzyme 2 (ACE2) is an IFN-stimulated gene and that SARS-CoV-2 could exploit species-specific IFN-driven upregulation of ACE2 to enhance infection [23].

One of the IFN-dependent TF that can bind several domains of the SARS-CoV-2 genome is SIX1. However, hypoxia and hypoxia-inducible factor (HIF) 1 α can increase SIX1 expression, which reciprocally stimulates HIF-1 α expression under both normoxic and hypoxic conditions, thereby creating a positive feedback regulatory loop [24]. Therefore, it is plausible that the hypoxic conditions in SARS-CoV-2 patients contribute to viral

activity by the activation of SIX1, which in turn binds to viral ORFs and genes.

Another factor that can bind SARS-CoV-2, HOXA9, plays an important role in glycoprotein C synthesis in the varicella zoster virus. Indeed, PBX/HOXA9 heterodimers bind the promoter region of this viral glycoprotein, and the HOXA9 homodimer binds the promoter of another viral protein which may upregulate C glycoprotein synthesis [25]. Moreover, we found a set of TF whose binding sites are present in SARS-CoV-2 but not in the bat CoV RaTG13, despite which they have a similarity of 96%.

As shown in our analysis of the 22 TF specific for SARS-CoV-2, we found GLIS2 and TFAP2A to likely be involved in ORF coding the spike protein. Moreover, we found TF involved in the virus life cycle, such as E2F1 and E2F4, MZF1 and TFAP2A.

In human herpesvirus 6, E2F1 can transcribe selected viral genes containing the E2F binding site in their promoters [26]. E2F4 permits hijacking of host cell-cycle regulation in order to create a favourable environment for the replication of the human parvovirus B19 [27]. MZF1 could potentially regulate the

TABLE 3. TF binding sites present exclusively in genome of SARS-CoV-2 isolated from humans and near starting site of each virus gene or ORF

TF name	TF ID	Start position	End position	Core match score	Matrix match score	Sequence
THRB (var. 2)	MA1575.1	5940	5958	0.964	0.858	TTGACCCTAAGTTGGACAA
GLIS2	MA0736.1	21633	21646	1	0.85	TACCCCTGCATAC
TFAP2A	MA0003.4	22207	22220	1	0.962	TCTCCCTCAGGGTT
PAX6	MA0069.1	17674	17687	1	0.878	ATCACGCATGATGT
MZF1	MA0056.2	29830	29842	1	0.952	GCTATCCCCATGT
KLF4	MA0039.4	11749	11760	1	0.885	ACTCCCCACCAA
SP9	MA1564.1	6188	6199	0.901	0.871	TACACACCTCT
SP9	MA1564.1	11746	11757	0.899	0.857	ACTACTCCACC
Nr1h3::Rxra	MA0494.1	241	259	0.962	0.877	TGTCCGGGTGTGACCGAAA
Nr1h3::Rxra	MA0494.1	20175	20193	0.962	0.904	TGTCCACAAATTACCTGAA
E2F1	MA0024.3	671	682	1	0.912	TACGGCGCGCAT
EGR1	MA0162.4	11747	11760	1	0.901	CTACTCCACCCAA
KLF11	MA1512.1	6188	6198	0.933	0.912	TACACACCTC
TBXT	MA0009.2	2506	2521	0.946	0.883	TCCCACAGAAGTGTTA
E2F6	MA0471.2	23604	23616	1	0.883	TCGGCGGGCACG
E2F4	MA0470.2	670	683	1	0.873	TTACGGCGCCGATC
KLF16	MA0741.1	6188	6198	0.908	0.9	TACACACCTC
Ebf2	MA1604.1	6662	6674	0.913	0.908	GTCCCTTGGGATA
ZBTB14	MA1650.1	6358	6369	1	0.855	GGACGCGCAGGG
EGR4	MA0733.1	11747	11762	0.893	0.882	CTACTCCACCCCAAGA
SP8	MA0747.1	6188	6199	0.96	0.892	TACACACCTCT
SP8	MA0747.1	11746	11757	0.964	0.876	ACTACTCCACCC
KLF5	MA0599.1	11750	11759	0.999	0.888	CTCCACCCCA
EBF3	MA1637.1	6662	6674	0.912	0.918	GTCCCTTGGGATA
KLF9	MA1107.2	6186	6201	1	0.886	ACTACACCCCTCTTT

We considered a range of ± 1000 bases from starting sites. Abbreviations: ORF, open reading frame; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; TF, transcription factor.

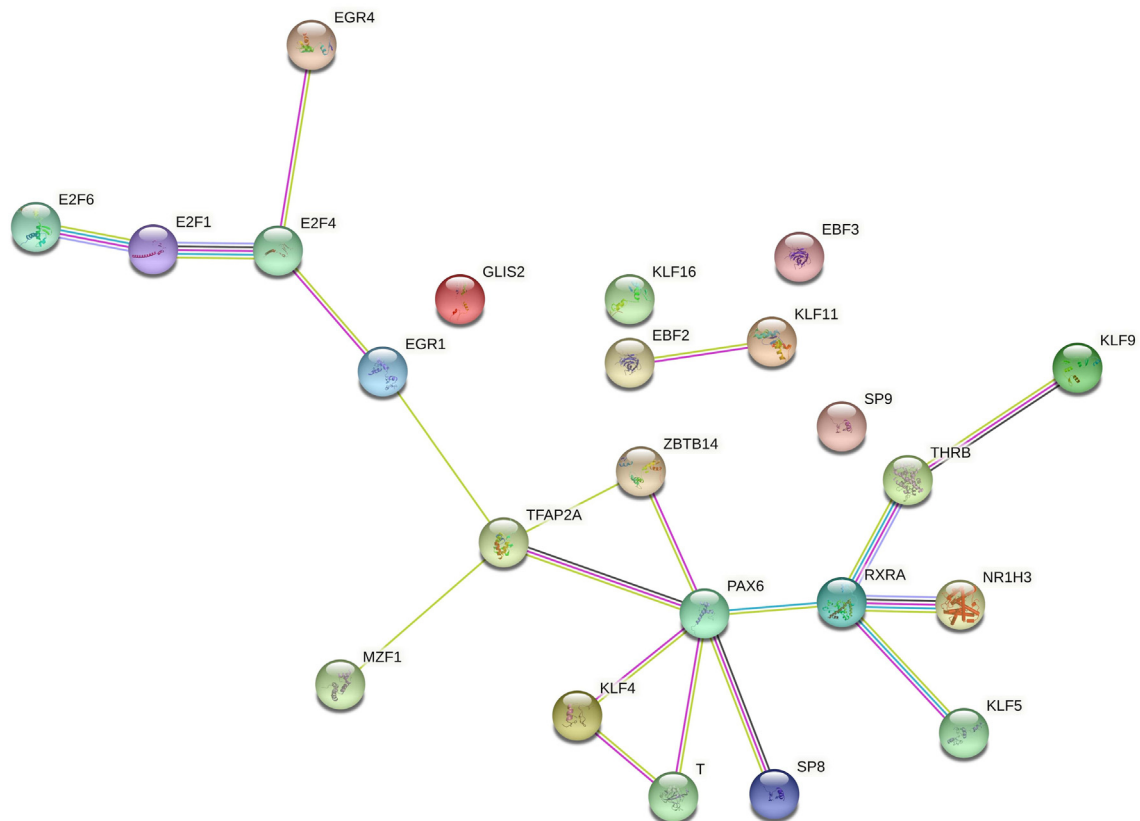


FIG. 4. Network analysis of transcription factors (TF) exclusively binding genomic region of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) isolated from humans. Twenty-two TF formed a significant network ($p = 8.27 \times 10^{-13}$) showing their involvement in retinoic acid signalling through direct binding of retinoic acid receptor RXR- α gene by NR1H3, KLF5, THR3 and PAX6 TF and indirect binding of remaining TF.

transcriptional programme associated with lethal infection related to H1N1 Influenza [28], whereas TFAP2A interacts with inducible viral and cellular enhancer elements to regulate transcription of selected genes [29]. Interestingly, the induction of TFAP2A promoter activity by NR2F2 was potentiated by the nuclear hormone receptors retinoic acid receptor alpha (RARA) and retinoid X receptor alpha (RXRA) [30]. In fact, as shown by our network analysis revealing TFBS to be present exclusively in human SARS-CoV-2, the strong involvement of retinoic acid signalling emerges. Most parts of the TFBS specific to the SARS-CoV-2 sequence are connected via the RXRA biological pathway (Fig. 4).

These data support the endogenous retinoic acid theory and the involvement of the retinoic acid metabolism in the COVID-19 cytokine storm [31,32]. Intriguingly, such metabolism is directly linked to the IFN response [31], therefore also explaining the involvement of the other TF identified by our enrichment analysis. In addition, our functional enrichment analysis showed several TF involved in the regulation of transcription by RNA polymerase II, suggesting a process similar to that of HIV regarding the recruiting of RNA polymerase II and TF machinery of the host to initiate viral transcription [33–35].

Finally, our analysis revealed that the binding site of the SPEDEF TF is present exclusively in the Brazilian SARS-CoV-2 variant and that this sequence, compared with the other, is enriched for GRHL1 and TFCP2 TF. The SPEDEF host TF has been found to be important in sustaining hepatitis C virus infection [36]. Moreover, it plays a critical role in regulating a transcriptional network mediating the mucus hyperproduction associated with chronic pulmonary disorders [37]. Therefore, the creation of this new TFBS in the Brazilian SARS-CoV-2 variant may explain why this virus variant results in a higher-severity respiratory syndrome. TFCP2 and GRHL1 TF are induced by IFN. In particular, TFCP2 is a factor that binds to and stimulates transcription from the viral SV40 major late promoter [38] and is involved in the regulation of HIV long terminal repeat [39]. Moreover, TFCP2 is involved in the regulation of interleukin-4 [38], a cytokine that upregulates mucus gene expression and mucus cell hypersecretion [40]. Remarkably, TFCP2 has also been identified as a key factor in the T-cell proliferative response [41]. It is interesting to note that both SPEDEF and TFCP2 are two host TF that can regulate the mucus secretion. Therefore, enrichment in the TFCP2 and GRHL1 TF host TFBS may be important for the severity of the respiratory syndrome associated with the Brazilian SARS-CoV-2 variant.

Our study has the limitation of having identified host TF interacting with SARS-CoV-2 only via an *in silico* analysis. However, the activation of TF depends on many parameters, in particular the activation state of the cell studied. A direct

molecular biological analysis of the interaction between these TF and their putative target viral sequence is needed. A new technique involving the attachment of an enzyme called a biotin ligase to the replicase complex could be used to validate the TF that come into contact with the SARS-CoV-2 replicase complex [9]. Moreover, specific SARS-CoV-2 TFBS could be cut out to prevent the binding of the specific TF and to evaluate the impact that it has on viral replication.

Conclusions

Our data suggest that SARS-CoV-2 may exploit the host TF involved in IFN, retinoic acid signalling and regulation of transcription by RNA polymerase II, thus facilitating its own replication cycle. This mechanism, by competition, may steal the human TF involved in these processes—a mechanism which has motivated researchers to produce a convincing body of evidence demonstrating that SARS-CoV-2 can disrupt the IFN-I signalling in host cells [32] and explaining the mechanism of SARS retinoic acid depletion syndrome leading to a cytokine storm [31].

These data strengthen our knowledge of the transcription and replicative activity of the virus, help to understand the mechanisms of interaction between SARS-CoV-2 and host, act as a starting point for further in-depth studies and could pave the way for new targets for drug design.

Conflict of interest

None declared.

Acknowledgements

Supported by University of Bari Aldo Moro and the Italian Ministry of Health (AIM-181005 to A. Stasi).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.nmni.2021.100853>.

References

- [1] Rabaan AA, Al-Ahmed SH, Haque S, Sah R, Tiwari R, Malik YS, et al. SARS-CoV-2, SARS-CoV, and MERS-CoV: a comparative overview. *Le Infez Med* 2020;2:174–84.

- [2] Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol* 2015;1282:1–23. https://doi.org/10.1007/978-1-4939-2438-7_1.
- [3] Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: structural genomics approach. *Biochim Biophys Acta Mol Basis Dis* 2020;1866:165878. <https://doi.org/10.1016/j.bbadis.2020.165878>.
- [4] Fanelli V, Fiorentino M, Cantaluppi V, Gesualdo L, Stallone G, Ronco C, et al. Acute kidney injury in SARS-CoV-2 infected patients. *Crit Care* 2020;24:155. <https://doi.org/10.1186/s13054-020-02872-z>.
- [5] Frutos R, Serra-Cobo J, Chen T, Devaux CA. COVID-19: time to exonerate the pangolin from the transmission of SARS-CoV-2 to humans. *Infect Genet Evol* 2020;84:104493. <https://doi.org/10.1016/j.meegid.2020.104493>.
- [6] Lai MM. Cellular factors in the transcription and replication of viral RNA genomes: a parallel to DNA-dependent RNA transcription. *Virology* 1998;244:1–12. <https://doi.org/10.1006/viro.1998.9098>.
- [7] Enjuanes L, Almázán F, Sola I, Zúñiga S. Biochemical aspects of coronavirus replication and virus–host interaction. *Annu Rev Microbiol* 2006;60:211–30. <https://doi.org/10.1146/annurev.micro.60.080805.142157>.
- [8] van Hemert MJ, van den Worm SHE, Knoop K, Mommaas AM, Gorbelenya AE, Snijder EJ. SARS-coronavirus replication/transcription complexes are membrane-protected and need a host factor for activity in vitro. *PLoS Pathog* 2008;4:e1000054. <https://doi.org/10.1371/journal.ppat.1000054>.
- [9] V'kovski P, Gerber M, Kelly J, Pfaender S, Ebert N, Braga Lagache S, et al. Determination of host proteins composing the microenvironment of coronavirus replicase complexes by proximity-labeling. *Elife* 2019;8. <https://doi.org/10.7554/eLife.42037>.
- [10] Gearing LJ, Cumming HE, Chapman R, Finkel AM, Woodhouse IB, Luu K, et al. CiiDER: a tool for predicting and analysing transcription factor binding sites. *PLoS One* 2019;14:e0215495. <https://doi.org/10.1371/journal.pone.0215495>.
- [11] Kel AE. MATCH™: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 2003;31:3576–9. <https://doi.org/10.1093/nar/gkg585>.
- [12] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;32:D91–4. <https://doi.org/10.1093/nar/gkh012>.
- [13] Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* 2018;46:D260–6. <https://doi.org/10.1093/nar/gkx1126>.
- [14] Boeva V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front Genet* 2016;7:24. <https://doi.org/10.3389/fgene.2016.00024>.
- [15] Cheon H, Holvey-Bates EG, Schoggins JW, Forster S, Hertzog P, Imanaka N, et al. IFN β -dependent increases in STAT1, STAT2, and IRF9 mediate resistance to viruses and DNA damage. *EMBO J* 2013;32:2751–63. <https://doi.org/10.1038/emboj.2013.203>.
- [16] Borden EC, Sen GC, Uze G, Silverman RH, Ransohoff RM, Foster GR, et al. Interferons at age 50: past, current and future impact on biomedicine. *Nat Rev Drug Discov* 2007;6:975–90. <https://doi.org/10.1038/nrd2422>.
- [17] Galán C, Sola I, Nogales A, Thomas B, Akoulitchev A, Enjuanes L, et al. Host cell proteins interacting with the 3' end of TGEV coronavirus genome influence virus replication. *Virology* 2009;391:304–14. <https://doi.org/10.1016/j.virol.2009.06.006>.
- [18] Emmott E, Munday D, Bickerton E, Britton P, Rodgers MA, Whitehouse A, et al. The cellular interactome of the coronavirus infectious bronchitis virus nucleocapsid protein and functional implications for virus biology. *J Virol* 2013;87:9486–500. <https://doi.org/10.1128/JVI.00321-13>.
- [19] Kropp KA, Angulo A, Ghazal P. Viral enhancer mimicry of host innate-immune promoters. *PLoS Pathog* 2014;10:e1003804. <https://doi.org/10.1371/journal.ppat.1003804>.
- [20] Mutthi P, Theerawatanasirikul S, Roytrakul S, Paemanee A, Lekcharoensuk C, Hansoongnern P, et al. Interferon gamma induces cellular protein alteration and increases replication of porcine circovirus type 2 in PK-15 cells. *Arch Virol* 2018;163:2947–57. <https://doi.org/10.1007/s00705-018-3944-1>.
- [21] Makvandi-Nejad S. Human immunodeficiency virus (HIV). British society for immunology. n.d. Available at: <https://www.immunology.org/public-information/bitesized-immunology/pathogens-and-disease/human-immunodeficiency-virus-hiv>.
- [22] Narayanan K, Ramirez SI, Lokugamage KG, Makino S. Coronavirus nonstructural protein 1: common and distinct functions in the regulation of host and viral gene expression. *Virus Res* 2015;202:89–100. <https://doi.org/10.1016/j.virusres.2014.11.019>.
- [23] Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, et al. SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. *Cell* 2020;181:1016–35. <https://doi.org/10.1016/j.cell.2020.04.035>.
- [24] Xu H, Zhang Y, Peña MM, Pirisi L, Creek KE. Six1 promotes colorectal cancer growth and metastasis by stimulating angiogenesis and recruiting tumor-associated macrophages. *Carcinogenesis* 2017;38:281–92. <https://doi.org/10.1093/carcin/bgw121>.
- [25] Storlie J, Jackson W, Hutchinson J, Grose C. Delayed biosynthesis of varicella-zoster virus glycoprotein C: upregulation by hexamethylene bisacetamide and retinoic acid treatment of infected cells. *J Virol* 2006;80:9544–56. <https://doi.org/10.1128/JVI.00668-06>.
- [26] Sharon E, Volchek L, Frenkel N. Human herpesvirus 6 (HHV-6) alters E2F1/Rb pathways and utilizes the E2F1 transcription factor to express viral genes. *Proc Natl Acad Sci U S A* 2014;111:451–6. <https://doi.org/10.1073/pnas.1308854110>.
- [27] Wan Z, Zhi N, Wong S, Keyvanfar K, Liu D, Raghavachari N, et al. Human parvovirus B19 causes cell cycle arrest of human erythroid progenitors via deregulation of the E2F family of transcription factors. *J Clin Invest* 2010;120:3530–44. <https://doi.org/10.1172/JCI41805>.
- [28] Josset L, Belser JA, Pantin-Jackwood MJ, Chang JH, Chang ST, Belisle SE, et al. Implication of inflammatory macrophages, nuclear receptors, and interferon regulatory factors in increased virulence of pandemic 2009 H1N1 influenza A virus after host adaptation. *J Virol* 2012;86:7192–206. <https://doi.org/10.1128/jvi.00563-12>.
- [29] Williams T, Admon A, Lüscher B, Tjian R. Cloning and expression of AP-2, a cell-type-specific transcription factor that activates inducible enhancer elements. *Genes Dev* 1988;2(12A):1557–69. <https://doi.org/10.1101/gad.2.12a.1557>.
- [30] Hubert MA, Sherritt SL, Bachurski CJ, Handwerker S. Involvement of transcription factor NR2F2 in human trophoblast differentiation. *PLoS One* 2010;5:e9417. <https://doi.org/10.1371/journal.pone.0009417>.
- [31] Sarohan AR. COVID-19: endogenous retinoic acid theory and retinoic acid depletion syndrome. *Med Hypotheses* 2020;144:110250. <https://doi.org/10.1016/j.mehy.2020.110250>.
- [32] Trasino SE. A role for retinoids in the treatment of COVID-19? *Clin Exp Pharmacol Physiol* 2020;47:1765–7. <https://doi.org/10.1111/1440-1681.13354>.
- [33] Ott M, Geyer M, Zhou Q. The control of HIV transcription: keeping RNA polymerase II on track. *Cell Host Microbe* 2011;10:426–35. <https://doi.org/10.1016/j.chom.2011.11.002>.
- [34] Roebuck KA, Saifuddin M. Regulation of HIV-1 transcription. *Gene Expr* 1999;8(2):67–84.
- [35] Ne E, Palstra RJ, Mahmoudi T. Transcription: insights from the HIV-1 promoter. *Int Rev Cell Mol Biol* 2018;335:191–243. <https://doi.org/10.1016/bs.ircmb.2017.07.011>.

- [36] Li Q, Brass AL, Ng A, Hu Z, Xavier RJ, Liang TJ, et al. A genome-wide genetic screen for host factors required for hepatitis C virus propagation. *Proc Natl Acad Sci U S A* 2009;106:16410–5. <https://doi.org/10.1073/pnas.0907439106>.
- [37] Chen G, Korfhagen TR, Xu Y, Kitzmiller J, Wert SE, Maeda Y, et al. SPDEF is required for mouse pulmonary goblet cell differentiation and regulates a network of genes associated with mucus production. *J Clin Invest* 2009;119:2914–24. <https://doi.org/10.1172/JCI39731>.
- [38] Ylisastigui L, Kaur R, Johnson H, Volker J, He G, Hansen U, et al. Mitogen-activated protein kinases regulate LSF occupancy at the human immunodeficiency virus type 1 promoter. *J Virol* 2005;79:5952–62. <https://doi.org/10.1128/jvi.79.10.5952-5962.2005>.
- [39] Shah S, Alexaki A, Pirrone V, Dahiya S, Nonnemacher MR, Wigdahl B. Functional properties of the HIV-1 long terminal repeat containing single-nucleotide polymorphisms in Sp site III and CCAAT/enhancer binding protein site I. *Virol J* 2014. <https://doi.org/10.1186/1743-422X-11-92>.
- [40] Khan MA, Khan ZA, Charles M, Pratap P, Naeem A, Siddiqui Z, et al. Cytokine storm and mucus hypersecretion in COVID-19: review of mechanisms. *J Inflamm Res* 2021;14:175–89. <https://doi.org/10.2147/jir.s271292>.
- [41] Kokoszynska K, Ostrowski J, Rychlewski L, Wyrwicz LS. The fold recognition of CP2 transcription factors gives new insights into the function and evolution of tumor suppressor protein p53. *Cell Cycle* 2008. <https://doi.org/10.4161/cc.7.18.6680>.