A SARS-CoV-2 host infection model network based on genomic human Transcription Factors (TFs) depletion

Massimiliano Chetta a,*, Alessandra Rosati b, Liberato Marzullo b, Marina Tarsitano a, Nenad Bukvic c

a Ospedale Antonio Cardarelli, C.O. di Genetica Medica e di Laboratorio, A.O.R.N. Cardarelli, Medical Genetics Laboratory, Building Y, Naples, Italy
b Department of Medicine, Surgery and Dentistry 'Schola Medica Salernitana', University of Salerno, Baronissi, 84081, Italy
c Azienda Ospedaliero Universitaria Consorziale Policlinico di Bari, Piazza Giulio Cesare 11, Bari, Italy

ARTICLE INFO

Keywords:
Bioinformatics
Genetics
Infectious disease
Virology
SARS-CoV-2 strain
Transcriptional factors depletion
Haploinsufficiency
Interactions within host genome

ABSTRACT

In December 2019 a new beta-coronavirus was isolated and characterized by sequencing samples from pneumonia patients in Wuhan, Hubei Province, China. Coronaviruses are positive-sense RNA viruses widely distributed among different animal species and humans in which they cause respiratory, enteric, liver and neurological symptomatology. Six species of coronavirus have been described (HCoV-229E, HCoV-OC43, HCoV-NL63 and HCoV-HKU1) that cause cold-like symptoms in immunocompetent or immunocompromised subjects and two strains of sometimes fatal zoonotic origin that cause severe acute respiratory syndrome (SARS-CoV and MERS-CoV). The SARS-CoV-2 strain is the emerging seventh member of the coronavirus family, which is actually determining a global emergency.

In silico analysis is a promising approach for understanding biological events in complex diseases and due to serious worldwide emergency and serious threat to global health, it is extremely important to use bioinformatics methods able to study an emerging pathogen like SARS-CoV-2.

Herein, we report on in silico comparative analysis between complete genome of SARS-CoV, MERS-CoV, HCoV-OC43 and SARS-CoV-2 strains, to identify the occurrence of specific conserved motifs on viral genomic sequences which should be able to bind and therefore induce a subtraction of host's Transcription Factors (TFs) which lead to a depletion, an effect comparable to haploinsufficiency (a genetic dominant condition in which a single copy of wild-type allele at a locus, in heterozygous combination with a variant allele, is insufficient to produce the correct quantity of transcript and, therefore, of protein, for a correct standard phenotypic expression).

In this competitive scenario, virus versus host, the proposed in silico protocol identified the TFs same as the distribution of TFBSs (Transcription Factor Binding Sites) on analyzed viral strains, potentially able to influence genes and pathways with biological functions confirming that this approach could brings useful insights regarding SARS-CoV-2. According to our results obtained by this in silico approach it is possible to hypothesize that TF-binding motifs could be of help in the explanation of the complex and heterogeneous clinical presentation in SARS-CoV-2 and subsequently predict possible interactions regarding metabolic pathways, and drug or target relationships.

1. Introduction

A novel coronavirus SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), responsible for the recent pandemic epidemic is a positive-sense RNA virus (ssRNA +), belonging to the Nidovirales order and Coronaviridae family, defined by a helical symmetry core and characteristic crown morphology [1]. Coronaviruses are genetically classified into four major genera: Alphacoronavirus, Betacoronavirus, correlated with mammal infections, and Gammacoronavirus, same as Deltacoronavirus that predominantly infects birds [2]. SARS-CoV-2 is a seventh strain of six different kinds of human CoVs previously identified: HCoV-NL63 and HCoV229E, that belong to the Alphacoronavirus genus; HCoV-OC43, HCoVHKU1, SARS-CoV (Severe Acute Respiratory Syndrome Coronavirus), and MERS-CoV (Middle East Respiratory syndrome Coronavirus), which belong to the Betacoronavirus group [3].

* Corresponding author.
E-mail address: mchetta@unisa.it (M. Chetta).

https://doi.org/10.1016/j.heliyon.2020.e05010
Received 5 May 2020; Received in revised form 30 July 2020; Accepted 17 September 2020
2045-8440/© 2020 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
The genome of SARS-CoV-2, is approximately 30,000 bases in length and is organized in 13–15 open reading frames (ORF) encoding a total of 27 proteins [2].

Like all coronaviruses the genome of SARS-CoV-2 encodes a complex of non-structural proteins polymeric RNA function (RNA-dependent RNA polymerase) and four main structural proteins: spike (S), membrane (M), envelope (E) and nucleocapsid (N) [2]. Additionally some other non-structural proteins are encoded in different numbers with uncertain functions. S glycoprotein, exposed on the virus surface, is the largest structural protein encoded by coronaviruses and is responsible for the binding and fusion of the virus with the host cell's membrane receptor. The protein S is encoded by an ORF with a highly variable nucleotide sequence that induces the main cell-mediated immune response. The host tropism of the coronaviruses is mainly determined by this protein which is composed of two portions: the N terminal end (S1) which performs a binding function to the receptor and the C-terminal end (S2) which performs the functions of fusion and entry of the virus into the cell [2].

The receptor binding domain (RBD) is a stretch of about 200 amino acids, located at the S1 end; it has the function of binding to the specific receptor on the surface of the host cell and represents the main immunostimulant epitope of the S protein [4]. Although, numerous variations have been identified in the sequence of SARS-CoV-2, it remains quite similar to SARS-CoV (about 79%) and MERS-CoV (about 50%) [2].

As expected, the greatest number of mutations take place in the region RBD and cause the recognition of different molecules on the surface of the host cell as a receptor. Human ACE2, expressed on the surface of the cells from respiratory and gastrointestinal tract, has been identified as the main cellular receptor for SARS-CoV-2 [5].

The growing evidence of the high adaptability of coronaviruses to new host animals in ecological niches (mammals, including humans, and birds) and the remarkable frequency of recombination, with a high mutation rate, required the use of tools that can provide quick information on the evolution of these viruses. Furthermore, the complex interplay between viruses and the host transcription machinery used by viruses for their own benefits, together with RNA viral replication, seems of fundamental importance.

New evidence confirmed that many positive and negative single-strand RNA viruses, whose primary replication site is cytoplasmic, are able to sequester nuclear proteins or TF to facilitate the replication process and alter the function of host cells. One of the mechanisms by which viruses can achieve this result is the disruption of the nucleus-cytoplasmic traffic and induce a spatial redistribution of proteins from the nucleus to the cytoplasm.

Several viruses (picornaviruses as well as VSV - vesicular stomatitis virus), use this kind of trafficking alteration to redistribute proteins to the cytoplasm, increase their interaction with the Internal Ribosome Entry Site (IRES) and facilitate translation of the virus polyprotein [6].

Recently was demonstrated that alphavirus transcripts are able to avidly bind cellular HuR protein through high-affinity binding sites conserved in their 3’-UTR. These RNA viruses, thanks to their ability to produce abundant amounts of viral RNA, can use this “sponge effect” as part of their strategy to modify cellular gene expression and interfere with cellular processes [7].

Among the emerging mechanisms of possible interactions between virus and host cell, it has been shown that many viruses (i.e. coronavirus SARS, influenza virus A, alpha and gamma herpesvirus) are able to increase the degradation of the host mRNA through interaction with specific viral proteins. These proteins can influence the cytoplasmic degradation of cytoplasmic mRNA by inducing widespread down-regulation of host gene expression. This mechanism, known as “host shutdown”, allows viruses to rapidly reduce the gene expression of the host cell and attenuate the immune response same as to recover proteins needed for viral replication [8].

Furthermore, it must be considered that although human cells contain the same genome, they can differ in the expression of genes. It is known that gene expression is subject to multiple effects both spatial and temporal and does not depend exclusively on the information contained in the coding sequence of the DNA. It has recently been shown that viruses by subtracting parts of the machinery of host cells can induce epigenetic and epitranscriptomic modifications of gene expression.

On the bases of these observations, we propose the hypothesis that the virus infection could be capable of inducing a subtraction of host Transcriptional Factors (TFs) which lead to a depletion, an effect comparable to haploinsufficiency (a genetic dominant condition in which a single copy of wild-type allele at a locus, in heterozygous combination with a variant allele, is insufficient to produce the correct quantity of transcript and, therefore, of protein, for a correct standard phenotypic expression). That means that RNA viruses, could be able to used this ability as extra opportunity of their strategy to usurp cellular gene expression and interfere with target cellular processes [7, 9]. In this competitive scenario, virus versus host, the analysis pipeline was performed using different bioinformatics friendly tools (available online free of charge) on the complete genome of SARS-CoV, MERS-CoV, HCoV-OC43 and SARS-CoV-2 strains, to identify the occurrence of specific conserved motifs capable to bind human TFs and subsequently predict their possible interplay.

2. Materials and methods

The analysis pipeline was performed using different bioinformatics tools available online and consists of four main steps:

1) Analysis of complete strains of SARS-CoV, MERS-CoV, HCoV-OC43 and SARS-CoV-2 to discover conserved motifs on RNA sequences. The analysis on *.FASTA sequences was performed using MEME (Multiple EM For Motif Elicitation). By default MEME chooses the width and number of occurrences of each motif automatically in order to minimize the ‘E-value’ of the motif, the probability of finding an equally well-conserved pattern in random sequences. Only motif widths between 6 and 50 are considered [10, 11].

2) All the obtained motifs were used as query for Tomtom (http://mem e-suite.org/doc/tomtom.html), another tool of MEME suite that compared the newly identified motifs against a database of known motifs (i.e., JASPAR). JASPAR CORE is a database that contained a curated and non-redundant set of open data access collections of experimentally discovered and proven TFs binding sites. Tomtom ranked the motifs in the database and produced an alignment for each significant match searching one or more query motifs against one or more databases of human target motifs (and their reverse complements when applicable). The report for each query was a list of target motifs, ranked by p-value and for each match an E-value and a q-value has been produced. The q-value is the minimal false discovery rate at which the observed similarity would be considered significant. Tomtom estimated q-values from all the match p-values using the Benjamini and Hochberg method. By default, significance was measured by q-value of the match. A list of Human TFs that contained the common conserved domain were obtained for all motif’s queries. All the human TFs were reported in Figure 1 [12,13].

3) Comparison between the all TFs obtained from the MEME output to select unique TFs for the SARS-CoV-2 strain.

4) Analysis on STITCH (Search Tool for Interactions of Chemicals) - http://stitch.embl.de/), a bioinformatics tool capable of providing integrated information about possible interactions regarding metabolic pathways, crystal structures, binding experiments, and drug or target relationships (Figure 1) [14].

3. Results and discussion

The original analytical workflow was developed to screen possible conserved motifs on SARS-CoV-2 able to subtract human genome TFs and subsequently modify the regulatory host scale networks. The result of the
in silico analysis permitted the identification of twenty unique TFs specific for SARS-CoV-2.

Using STITCH analysis, fifteen of those TFs were connected in a gene regulatory knowledge network, while five (HIC1 - HIC Zinc Finger Protein 528 and NR1I2 - Nuclear Receptor Subfamily 1 Group I Member 2) seems to be without known interactions (Figure 2, panel A).

The approach adopted by our in silico analysis allowed to predict that a reduction of some TFs directly related to diffuse parenchymal lung disease, beside well known roles in the immune system. In particular, GATA2 (GATA-Binding Protein 2) deficiency is reported in children and adults with severe pulmonary alveolar proteinosis and hematologic disorders [15], while NFATC3 (Nuclear Factors of Activated T Cells Cytoplasmic, Calcineurin-Dependent 3) alteration is involved in proliferation of human pulmonary fibroblasts after hypoxic stimulus [16], and, finally, STAT4 (Signal Transducer and Activator of Transcription 4) associated to protective function in patients with systemic sclerosis [17]. The synergic action of these TFs could further clarify the mechanism of lung complications triggered by the infection of SARS-CoV-2.

Furthermore, the analytical workflow permitted to search indications for possible treatment on the basis of interaction between chemical compounds and TFs by selecting the function of chemical display directly on STITCH website. In particular, Tacrolimus, which is directly correlated with NFATC3 (Figure 2, panel B), has been identified. Tacrolimus used in therapy of patients for whom immunosuppression is needed, gives a promising results also in those patients with severe connective tissue interstitial lung disease [18]. With the same approach also Berberine a quaternary ammonium salt from the protoberberine group of benzylisoquinoline alkaloids has been individualized. This alkaloid reduces a fibroproliferative lung disorder in Idiopathic Pulmonary Fibrosis (IPF) of unknown etiology by modulation of Transforming Growth Factor-β/1 (TGF-β1)-mediated SMAD (SMAD Family Member 2; Mothers Against Decapentaplegic Drosophila Homolog 2) and non-SMAD signaling cascades (Figure 2 panel C) [19]. A very similar mechanism of action involving Janus Kinase 1 (JAK 1) and JAK2 has been described for the Ruxolitinib an oral inhibitor of JAK protein currently under investigation for possible use in treatment of COVID-19. These observations could be tested in biological systems and could be useful in processes towards the clarification of novel therapeutic aspects.

Moreover (Figure 2, panel A), our analysis indicates a close connection between different TFs capable of inducing a modulation of immune response. In particular, a deficiency of Activating Transcription Factor 4 (ATF4) that during Endoplasmic Reticulum (ER) stress-mediated inflammation, increases expression of inflammatory cytokines, including interleukin-6 (IL-6) and monocyte chemoattractant protein-1 (MCP-1), by inducing expression of NACHT, LRR, and PYD domain-containing protein 1 (NLRP1), a core component of the inflammasome [20]. The modification of interleukin (IL-6) expression, a multifunctional pro-inflammatory and anti-inflammatory cytokine, in turn [21], reduces promptly and transiently the response to infections and in particular to severe pneumonia induced by bacterial (i.e. S. Pneumoniae) [22]. The role of IL-6 as an anti-inflammatory cytokine is given, also, by its inhibitory effect on TNF-α (Tumor Necrosis Factor Alpha) and IL-1, same as by the activation of IL-10 and IL-1Ra (Interleukin 1 Receptor Antagonist) [23]. Furthermore, ATF4 is also implicated in the saturated fatty acid (SFA)-induced IL-6 expression in macrophages. Attenuation of ATF4 in macrophages markedly inhibits SFA-induced IL-6 expression [24]. Similarly is conceivable an inhibition of Interferon Regulatory Factor-1 (IRF1), the founding member of the interferon regulatory factor family, a transcription factor that regulates a range of target genes involved in vivo response to viruses and bacteria, and NFATC3 (Nuclear factor of
activated T-cells, cytoplasmic 3) that plays a role in the induction of the IL-2 and expression of cytokine genes in T-cells [25, 26]. Through this approach CEBPG (CCAAT Enhancer Binding Protein Gamma) was identified too. This protein binds the promoter and the enhancer regions of target enhancer element PRE-I (Positive Regulatory Element-I) inducing the IL-4 gene expression, POU2F1 (POU Domain, Class 2, Transcription Factor 1) that modulates transcription of viral early genes immediately and STAT4 (Signal Transducer and Activator of Transcription 4) associated with a reduction of IFN-γ immunity and macrophage activity susceptibility [27, 28, 29].

It is particularly interesting that the modulation of the described TFs could be traced to the highest variability degree of IL1, IL6, IL10 and TNF-α related to the response against exogenous effects [30] and how these cytokine are part of three clusters with similar responses: first [TNF-a, IL1, IL10], second [IFN-c, IL2, IL4, IL8, and IL12], and third [IL6]. Association of these cytokines and subsequent cellular adaptive immune system regulation and secretion are involved in the appearance of cytokine storm [30, 31, 32].

The analysis was expanded to characterize other possible implications related to the deficiency of other TFs (Figure 2, panel A). In particular, those TFs with apparent unknown interactions (HIC1, DUX4, FOXQ1, ZNF582), are involved in embryonic development, cell cycle regulation, tissue-specific gene expression, cell signaling, and tumorigenesis. Consequently, if infection occurs during precocious weeks of pregnancy a decrease of these TFs could be traced to a possible implication in embryonic development and possible fetal malformations [33, 34, 35, 36].

DLX3 (Distal-Less Homeobox 3) and HOXC9 (Homeobox C9), which, respectively, determine reduced placental formation of blood vessels and

![Fig. 2. The scheme provides an overview of the TFs computational relationship analysis (based on STITCH tool) associated with each functional clusters and drugs (panel A). Panel B and C shows the paths related to tacrolimus and berberine respectively.](image-url)
increased pre and perinatal consequences [37, 38] could be added to these TFs. HIC1 is also correlated with allele accelerates polyp formation in adenomatous polyposis coli (APC) and in sporadic gastrointestinal (GI) cancers [39]. Of particular interest is NR112 (Nuclear Receptor Subfamily 1, Group I, Member 2), that when down-regulated, does not activate the transcription of multiple genes involved in the metabolism and secretion of potentially harmful xenobiotics (drugs and endogenous compounds). This down-regulation produces a multi drug resistance, specifically to rifampicin [40].

Interestingly, GATA2 (GATA Binding Protein 2) a transcription factor with an essential role in proliferation and differentiation of hematopoietic cells, as mentioned earlier, was identified among the TFs. An anomaly in GATA2 has been recently reported as the cause of familial syndromes with autosomal-dominant inheritance such as severe monocytopenia, NK and B lymphopenia, and absence of dendritic cells. Gata2 reduction also confers a selective advantage to EVII (Ectotrophic Viral Integration Site 1) expression that contributes to an acceleration of leukemogenesis [41, 42].

Another group of TFs seems to be involved in possible development of human malignancies (see Figure 2, panel A,B,C). In particular the translocations and deletions of JUN (AP-1 transcription factor) and JUND (JunDProto-oncogene, AP-1 Transcription Factor Subunit) related to deficiency of expression, is correlated to a development of predisposition of myelodysplastic syndrome and chronic myelomonocytic leukemia [43]. Besides these SMAD2 deficiency (SMAD Family Member 2) is related to an alteration of Tgfr1 signaling and induced development of colorectal cancer [44].

Finally, TBP (TATA-Box Binding Protein) a member of evolutionarily conserved proteins known as TAF (TBP Associated Factors), a co-activator that facilitates RNA polymerase II complex assembly at the beginning of transcription, could be considered as another possible target for the development of therapy. TBP, is a target of Ribavirin a nucleoside analogue that interferes with viral synthesis of guanosine triphosphate. Although treatment with Ribavirin has been demonstrated to have no real efficiency on clinical outcome of patients with SARS or MERS, the possible direct interaction on TBP could contribute to reconsideration as a useful drug when administered alone or in combination with other compounds for COVID-19 [45,46].

In regards to the possibilities of selective therapeutic treatments of COVID-19 (SARS-CoV-2 infection), it seems possible that expression of four TFs (FOXO1, HIC1, HOXC9, JUN), selected by our analysis, could be modulated by treatment with Pirfenidone (Figure 2, panel A). This drug reduces fibroblast proliferation, production of fibrosis-associated proteins and cytokines, as well as an increased biosynthesis and accumulation of extracellular matrix in response to cytokine growth factors such as Transforming Growth Factor - Beta (TGF-β) and Platelet - Derived Growth Factor (PDGF); used to treat idiopathic pulmonary fibrosis [47].

Finally, our analysis indicates that a possible deficiency of TBX21 (T-Box Transcription Factor 21) is associated with significant risk of Aspirin-induced asthma [48]. Therefore, at least in the early stages of the illness, it may be prudent to avoid use of this non-steroidal anti-inflammatory drug (NSAID).

4. Conclusion

In conclusion, the extended knowledge of TFs identified by this in silico approach with apparent deep translational relevance could provide an insight of the molecular aspects as they relate to the infection process, the dynamics of changes in the host's cellular system; clarifies the possible mechanism of elevated pulmonary tropism and, finally, indicate some clarifications that reflect on paving the way to the best therapeutic strategies. Even though further in vitro and in vivo analysis will be necessary to support the functional evidences derived from above reported observations of our in silico research, it seems reasonable to advise that all patients cured from SARS-CoV-2 infection, to undergo long term follow up for tumor susceptibility due to possible oncogenic potential of SARS-CoV-2. Same attention should be paid in regards to the potential pregnancy risks.

Genetic variation is the process, devoid of specific and predetermined purposes, which underlies the differentiation and, therefore, the evolution of viruses. The set of viral mutations that originate as a consequence of different molecular mechanisms, allow the virus adaptation to the environment that will act by selecting the most suitable characteristics [49, 50]. RNA viruses, with high predisposition to replication error mediated by RNA-polymerase or reverse transcriptase are under the high diversification process [51]. The introduction of errors in the virus genome, generally around leading or consensus sequences, determines an extremely heterogeneous population structure which, in many cases, follows the concept of “quasispecies” - the viral population within a host composed by a distribution of different mutants in dynamic balance [52, 53]. This population is subject to continuous processes of genetic variation, competition and adaptation, that favor subsequent selection. Surface proteins that bind to the cell receptor, even if the non-structural proteins and the regulatory regions of the viral genome can influence the tropism of virus for the host cell play the main role in this selection [54]. Considering the evidence as well as the public health emergencies such as COVID-19 (SARS-CoV-2), it is necessary to find new rapid analysis strategies, such as computational approaches which, in addition to providing indications on the mechanisms of action of the virus, can also provide evolutionary models and possible models of transmission following the disturbance of the host's cellular network.

Even though in silico analyses require in vitro and in vivo validations, undoubtedly could provide valid contribution to gain deeper understanding of COVID-19 (SARS-CoV-2) and could be of help in global effort to find out effective measures, designing a suitable vaccine and/or therapeutic strategy [55, 56, 57].

Declarations

Author contribution statement

Massimiliano Chetta: Conceived and designed the analysis; Analyzed and interpreted the data; Wrote the paper.

Alessandra Rosati: Analyzed and interpreted the data; Wrote the paper.

Liberato Marzullo: Analyzed and interpreted the data.

Marina Tarabulino: Contributed analysis tools or data.

Nenad Bukvic: Conceived and designed the analysis; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] A.R. Fehr, S. Perlman, Coronavirus: an overview of their replication and pathogenesis, Methods Mol. Biol. 1282 (2015) 1–23.

[2] A. Wu, Y. Peng, B. Huang, X. Ding, X. Wang, P. Niu, J. Meng, Z. Zhu, Z. Zhang, J. Wang, J. Sheng, L. Quan, Z. Xia, W. Tan, G. Cheng, T. Jiang, Genome composition and divergence of the novel coronavirus (2019-ncov) originating in China, Cell Host Microbe 27 (3) (2020 Mar 11) 325–328.

[3] K.G. Andersen, A. Rambaut, W.i. Lipkin, et al., The proximal origin of SARS-CoV-2, Nat. Med. (2020).
[51] S. Duffy, Why are RNA virus mutation rates so damn high? PLoS Biol. 16 (8) (2018 Aug 13), e3000003.

[52] S.L. Fishman, A.D. Branch, The quasispecies nature and biological implications of the hepatitis C virus, Infect. Genet. Evol. 9 (6) (2009 Dec) 1158–1167.

[53] E. Domingo, J. Sheldon, C. Perales, Viral quasispecies evolution, Microbiol. Mol. Biol. Rev. 76 (2) (2012 Jun) 159–216.

[54] M.S. Maginnis, Virus-receptor interactions: the key to cellular invasion, J. Mol. Biol. 430 (17) (2018 Aug 17) 2590–2611.

[55] E. Kim, G. Erdos, S. Huang, W.T. Kenniston, C.S. Balmert, Donahue Carey, V. Stalin Raj, W.M. Epperly, B.W. Klimstra, L.B. Haagmans, E. Korkmaz, L.D. Falo Jr., A. Gambotto, Microneedle array delivered recombinant coronavirus vaccines: immunogenicity and rapid translational development, EBioMedicine (2020).

[56] C. Chakraborty, A.R. Sharma, G. Sharma, M. Bhattacharya, S.S. Lee, SARS-CoV-2 causing pneumonia-associated respiratory disorder (COVID-19): diagnostic and proposed therapeutic options, Eur. Rev. Med. Pharmacol. Sci. 24 (7) (2020) 4016–4026.

[57] D.E. Gordon, et al., A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, Nature (22 April 2020).