1    **Transcriptional Difference between SARS-COV-2 and other Human Coronaviruses Revealed by Sub-**

2    **genomic RNA Profiling**

3    Lin Lv[#], Xiaoqing Xie[#], Qiyu Gong, Ru Feng, Xiaokui Guo, Bing Su, Lei Chen*

4

5    Shanghai Institute of Immunology, Shanghai Jiao Tong University School of Medicine, Shanghai 200025,

6    China

7

8    # These authors contributed equally.

9    * To whom correspondence should be addressed: Lei Chen (lei.chen@sjtu.edu.cn).

10

11

12    **Abstract**

13        SARS-COV-2 and all other coronaviruses express its 3 prime genes by forming sub-genomic RNA. As

14    the genome of these virus exist in RNA form, only by profiling the relative abundance of these sgRNAs, can

15    the viral transcriptome be revealed. Utilizing publically available meta-transcriptomic data generated from

16    patient samples, we were able to infer the viral transcriptome in vivo, which is distinct from the *in vitro* one

17    derived from cell culture. Inter-sample diversity was also observed and a sample specific transcript was

18    identified. By doing the same analysis to MERS and SARS data, we were able to compare the three in terms

19    of transcription. Among the differences, SARS-COV-2 has significantly elevated expression of the Spike gene,

20    which may contribute to its high transmissibility.

21

22    **Highlights**

23    1)    The *in vivo* transcriptome of SARS-CoV-2 revealed by sgRNA profiling, for 25 patient samples around

24         the globe.

25    2)    The Spike protein expression is an order of magnitude higher in SARS-CoV-2 than MERS-CoV or SARS-

26         CoV, possibly contributing to the virus' elevated transmissibility.

27    3)    The *in vivo* SARS-CoV-2 transcriptomes, as inferred from human patient data was distinct from the *in*

28         *vitro* one derived from cell line culture, all the accessory genes were up-regulated *in vivo*, suggesting

29         intricate expression regulation mechanism for the small viral genome.

30

**Introduction**

COVID-19 has reached global pandemic levels since March 2020, brought unprecedented devastation to human lives and to global economy as well [1]. Like other coronaviruses, the genome of SARS-COV-2, the virus responsible for COVID-19, exists in the form of a single positive strand of RNA. Upon entrance into host cells, the open reading frame (ORF) closest to 5' end can be readily translated from the genome, in reality this ORF is divided into two segments—ORF1a and ORF1ab, due to a ribosome skipping mechanism involved [2]. All other 3' ORFs, including ones coding for the structural proteins: S (Spike), E (Envelope), M (Membrane) and N (Nucleocapsid) and a number of accessory genes, can only be translated by first forming the so called sub-genomic RNAs or sgRNAs. These sgRNAs are formed by recombination of the 3' portion of the genome with a leader sequence of around 70 nt. The recombination is usually thought to be based on a repetitive sequence called Transcription Regulation Sequence (TRS), Fig 1A.

The transcriptome of the virus is a fundamental aspect of its biology. For example, knowing which accessory protein is expressed and in which strain can greatly help us combating it. Also in the 21 century, two previous episodes of coronavirus outbreaks had plagued us: the SARS (Severe Acute Respiratory Syndrome) outbreak in 2003 and MERS (Mid-Eastern Respiratory Syndrome) outbreak in 2012. These two outbreaks caused far fewer cases, 8071 and 2499 respectively, whereas the fatality rates were much higher, 9.6% for SARS and 37.0% for MERS, compared with a current estimate of 4.5% for SARS-COV-2 [1, 3, 4]. Studying the difference between the three can reveal valuable information on their different transmissibility and virulence. But it's currently lacking beyond clinical information or simple genome comparison due to the difficulty in handling these deadly viruses.

Sequencing has played essential role in identifying this particular virus and are still contributing to the diagnosis and strain typing to keep the spread and evolution of it under close monitoring [5, 6]. The vast majority of these sequencing data come in the form of meta-transcriptomics sequencing, i.e. the sequencing of the RNA of all organisms within a biological sample, mostly bronchoaveolar lavage fluid or BALF. Usually the viral reads within these data were extracted and assembled or used to call SNPs, so the underlying strain would be typed and compared to existing ones. Once sequenced, early annotation of the viral genome were made based on sequence homology [5, 6]. However, homology alone does not warrant expression of the underlying gene as these highly mutable RNA virus can have other sequence features altered that rendering ORFs that looked intact not expressed. Recently, direct profiling of viral RNAs were made using nanopore technology that revealed the transcriptome of the virus, these study all used isolated

61  virus strain to infect VERO cells to generate the sample for sequencing [7-9]. In this paper, we turned to the

62  readily available public data. By profiling the sgRNA population underlying the data using an alignment

63  break point analysis method, we were able to reconstruct viral transcriptome in vivo for a large number of

64  samples, and study their diversity. Doing the same to MERS and SARS, we were able to compare the

65  transcriptomes of these human coronaviruses.

66

67  **Results**

68      As of 2020/03/31, there are more than 2800 viral genome sequences submitted to GISAID. However

69  there are only 11 Bioprojects containing almost 60 samples with raw data deposited in NCBI SRA. After

70  filtering out samples that contained too few viral reads and adding ones for SARS-CoV and MERS-CoV, we

71  obtained 8 projects with 42 samples and 698 million Illumina reads in total, Supplementary Table 1. All the

72  SARS-Cov-2 samples came from human, while SARS and MERS ones came from mouse or cultured cells.

73      We developed an informatics pipeline to infer sgRNA profiles from these short reads samples. In short,

74  reads were first aligned to a combined reference of human (hg38) and either SARS-CoV-2 (GeneBank ID

75  MT121215.1), SARS (GeneBank ID NC_004718.3) or MERS (GeneBank ID NC_019843.3). The relative

76  abundance of the sgRNA species can then be inferred from alignment break point analysis. I.e. the reads

77  that derive from the 5' end of the sgRNAs, when mapped to whole genome reference, will have one end

78  mapped to the 5' leader and one end mapped far away from it, Fig 1A. See supplementary methods for

79  more details.

80      For SARS-CoV-2, there are 25 samples in total, between 2.99% and 99.97% of the reads were mapped

81  to SARS-CoV-2 (84.76% average, the large variation arose from different sequencing library construction

82  methods used). The sgRNA profiling identified 8 cannonical break points corresponding to 8 sgRNA species:

83  S, E, M, N, ORF3a, ORF6, ORF7a and ORF8, Fig1B. N is the most abundant at 28% of the canonical sgRNAs.

84  It's followed by ORF7a, M, ORF3a and S, all just above 10%. E is the lowest of the eight at about 1%. All of

85  them also have TRS except E. The break point for these ORFs were situated between 9 to 146 nt upstream

86  of the ORF start, and have no start codon (ATG) in-between. ORF7b and ORF10 are two ORFs that were

87  supported in a recent proteomics study on SARS-CoV-2. However, we did not see supporting break

88  points/sgRNA in our first scan-through. We then looked extra hard by lowing our threshold. The few extra

89  break points emerged were still very far away from those ORF start or has out-of-frame start codon between

90  the break point and the ORF start. Thus ORF7b and ORF10 were not supported by our data, this could be

91    due to lack of read depth or difference between *in vivo* and *in vitro* situations.

92    We also examined a few well supported break point without corresponding canonical ORF annotations

93    and found a novel ORF starting at 22614, encoding a 36 amino acid peptide. This break point was only

94    identified in one of the samples but reached high frequency in it. Interestingly, this novel break point has

95    an underlying TRS that's shifted several bases from the canonical one and all the samples contain this

96    shifted TRS. The underlying sample (SRA ID SRR11278164) has a genomic sequence differs from other

97    samples mostly at ORF1ab, suggesting replication/transcription machinery change rather than TRS

98    alteration contributed to this novel sgRNA. Although if this novel sgRNA or viral transcript was translated

99    or not remain to be seen.

100
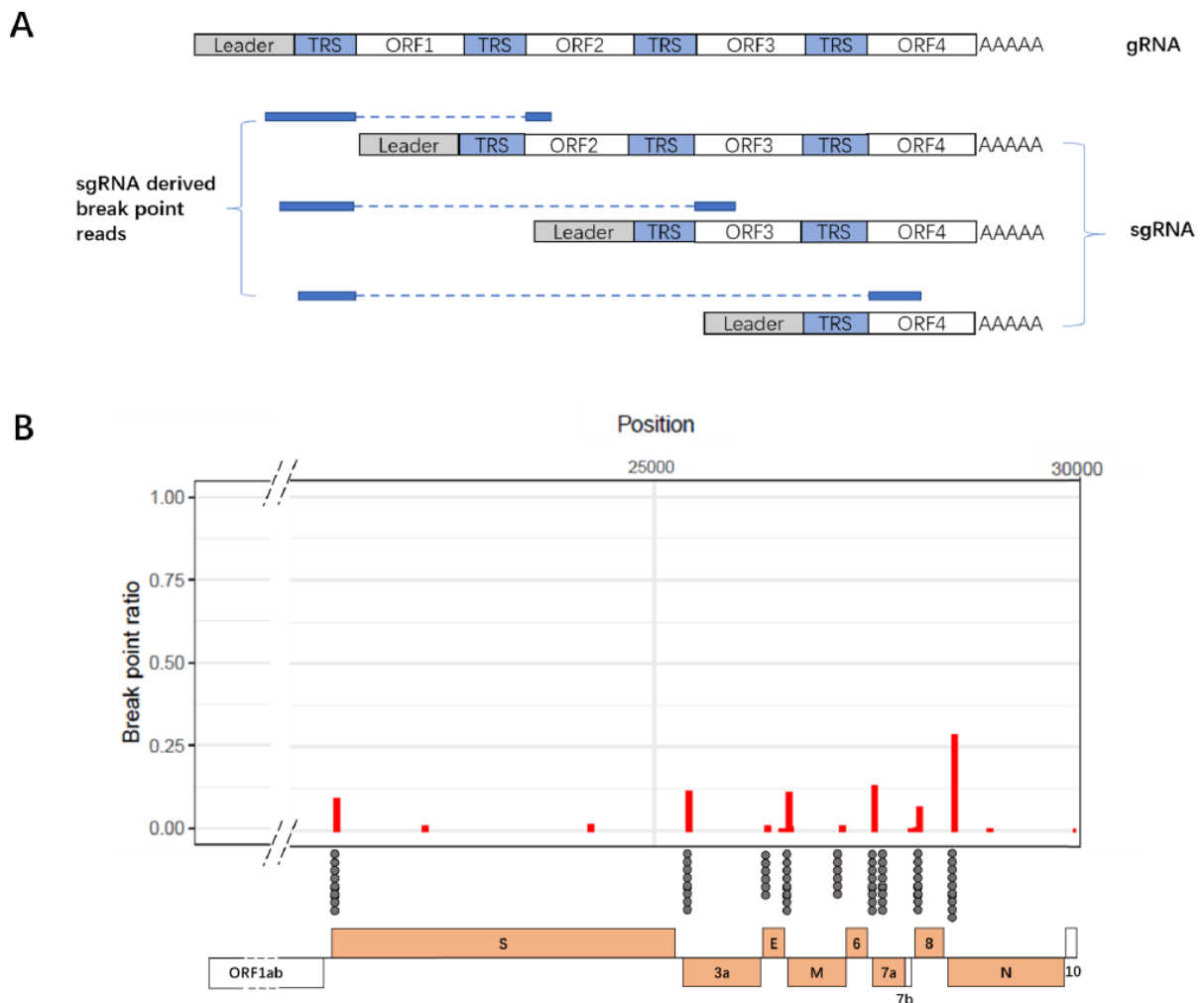


101

102   **Figure 1. The sgRNA profile for SARS-COV-2.** A) schematic showing of gRNA and sgRNAs for coronaviruses,

103   the gRNA translate the 5' ORF, while sgRNAs are formed by concatenation of variable lengths of the 3'

104   portion of gRNA with its leader, usually through a TRS (Transcription Regulation Sequence), all the 3' ORFs

105     are translated from their respective sgRNAs, as usually only one ORF is translated from a messenger RNA.

106     B) Top, the break point ratios; middle, the TRS sequences within the SARS-COV-2 genome, solid dots

107     indicate a matching base to leader TRS sequence while a hollowed one indicate a mismatch; bottom, the

108     schematic showing of the SARS-COV-2 annotation, the ORF supported by sgRNAs indicated by orange color.

109     For SARS, there are 10 samples in total, between 0.10% and 55.37% of reads were mapped to

110     SARS(14.59% average). The number of 3' ORFs annotated by homology alone is 12 while sgRNA profiling

111     suggest 8, Suppl. Fig 1. For MERS, there are 7 samples in total, between 0.04% and 96.21% of reads were

112     mapped to MERS (14.78% average). Homology alone suggest there is 9 3' ORFs in this genome, while sgRNA

113     profiling suggest 7, Suppl. Fig 2. This makes MERS the one with fewest genes amongst the three. When the

114     annotation of the 3 viruses were compared, both homology based and sgRNA based, SARS comes closest

115     to SARS-CoV-2, as expected. Also, sgRNA evidence suggest both express 8 3' ORFs, with an almost perfect

116     correspondence, with ORF8 from SARS-CoV-2 being the exception. It shares weak and partial homology

117     with SARS-CoV ORF8b, not the well express 8a, Fig 2A.

118     Since we have multiple samples for each of the three viruses, we can compare their transcriptome

119     across samples, Fig 2B. Note that after imposing a minimum break point read requirement, the number of

120     samples for the SARS-CoV-2 group was down to eight. The most striking different came from S. The relative

121     expression of S gene in SARS-Cov-2 is 14.8%, an order of magnitude higher than that in SARS and MERS,

122     1.6% and 1.25% respectively. We notice this difference is in accordance with the three's capability to spread

123     among hosts: MERS being the most deadly, has a R0 of 0.69, SARS-CoV-2 having the lowest fatality rate,

124     spread very rapidly and SARS comes in between in terms of both transmissibility and fatality rate [10-12].

125     SARS-CoV-2 and SARS-CoV both use ACE2 as the receptor for celluar invasion. Cryo-EM study suggested

126     that the SARS-CoV-2 has higher binding affinity to ACE2 [13]. Our study suggest yet another possibility. The

127     elevated S expression, may make an average SARS-COV-2 having more Spikes on surface or just assemble

128     more efficiently.

129     We also compared the recent in vitro transcriptome with our in vivo one, Fig 2B, green dashed lines.

130     An interesting observation is that all accessory genes were down-regulated in vitro. This could make
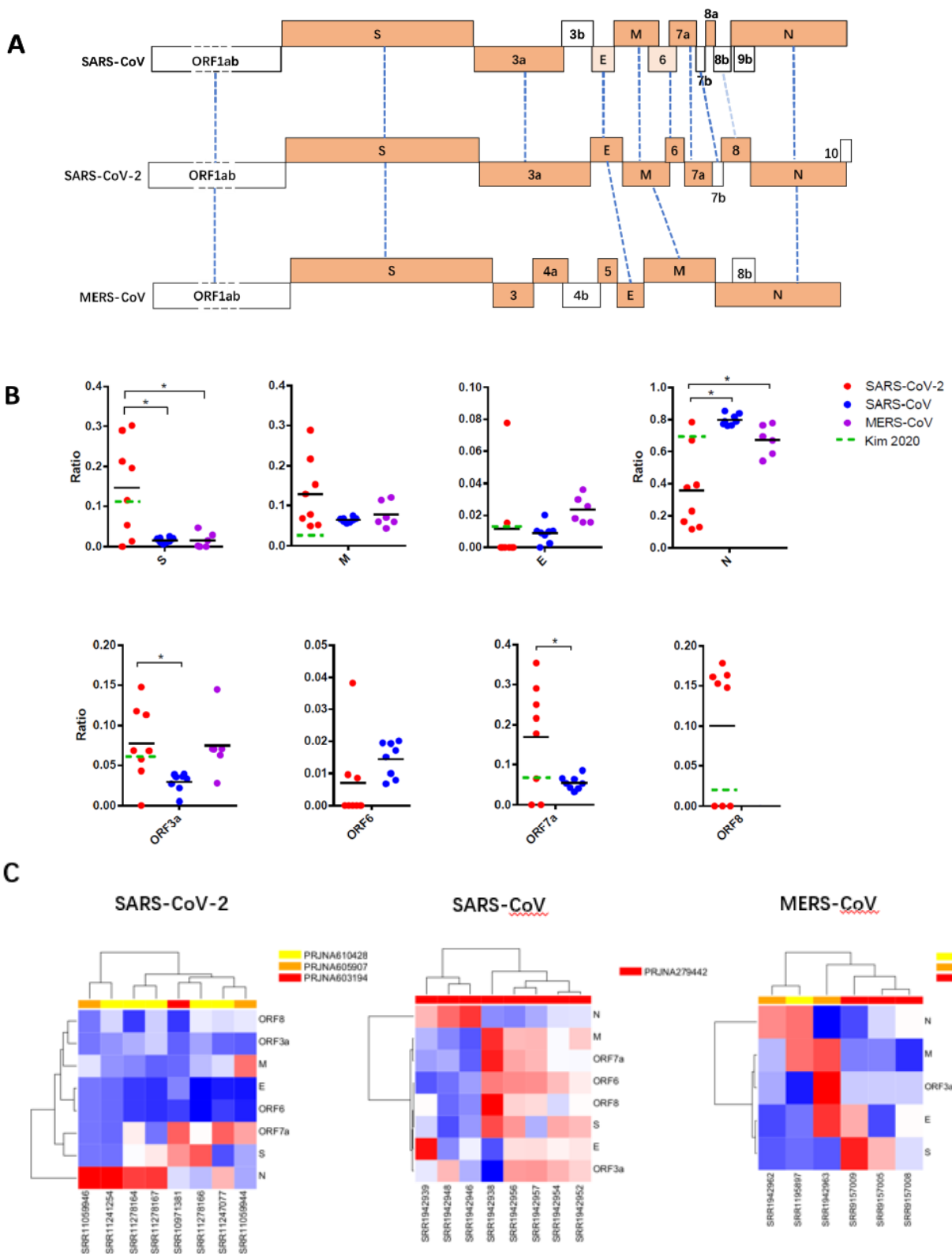
131     biological sense as in defenseless cultured cells, viruses devote more resource into reproduction, while to

132     cope with more host defense *in vivo*, accessory genes have to more important roles to play. This also

133     suggested there was intricate expression regulation mechanism at work here, even for this small 30 Kb

134     genome.

135



136

137

138    **Figure 2. Comparison between SARS-COV-2, SARS and MERS in terms of sgRNA profiles.** A) The schematic

139    showing of correspondence between the three viruses, the orange colored ORFs were supported by sgRNA

140     analysis; B) The sgRNA profiles of the 3 viruses compared; C) The heatmaps of the viral expression matrices,

141     showing clustering of samples as well as genes.

142

143     Lastly, we made clustered heatmaps for the viral transcriptome matrices, Fig 2C. The samples did not

144     cluster by Bioproject ID, suggesting low batch effect across a wide range of library processing methods,

145     validating this type of short reads break point profiling as relatively unbiased and the inferred transcriptome

146     can be compared. Clustering of the genes suggest that for SARS-CoV-2, the E gene and ORF6 gene were

147     highly correlated. Interestingly these two genes are of similar length, having similarly sized TRS and similar

148     sgRNA abundance.

149

150     **Discussion**

151     SARS-COV-2 were identified very early on by sequencing technology. The sequence data on this virus

152     were still generated in high volumes around the globe as of this paper. These sequencing data were so far

153     used mostly for strain typing the virus. This is an under-utilization of a valuable information. By developing

154     an informatics pipeline, we were able to infer the sgRNA populations from these data, essentially recover

155     the viral transcriptome. This present an added layer of information on top of the genome sequence. Our

156     results suggest that different strains of virus may even have a different repertoire of genes expressed,

157     highlighting the need for more in-depth case examination and individual patient care.

158     Spike protein received tremendous amount of research attention as it determines the host specificity

159     and receptor affinity thus transmissibility. Our result show that the SARS-CoV-2 Spike gene is highly

160     expressed in at least a large number of virus strains, on average an order of magnitude higher than SARS-

161     CoV or MERS-CoV. This may also contribute to the elevated transmissibility of this virus. Carefully controlled

162     electron microscopy comparison of these virion particles may be able to answer if the relative amount of

163     the Spike is really different

164     Finally, there are more than just viral information within this type of meta-transcriptomic data. Host

165     RNA and RNA from other microbes were also there, though not optimally captured. In a time of urgent

166     need, these valuable information should not be wasted, and could be made more useful if more

167     accompanying clinical information could be shared. Most GISAID entries for SARS-CoV-2 has a meta-

168     transcriptomic dataset behind it and currently GISAID entries far out-numbered the raw reads entries in say

169     SRA. Sharing these information will greatly help researchers study this virus and ultimately curb it.

170

## References

172   1.   **World Health Organization:Coronavirus disease 2019 (COVID-19) Situation Report** [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports]

174   2.   Straus SE: **Coronaviridae.** In *Fields Virology.* Edited by Knipe DM, Howley PM. Philadelphia: Lippincott Williams & Wilkins; 2013: 825-859

176   3.   Memish ZA, Perlman S, Van Kerkhove MD, Zumla A: **Middle East respiratory syndrome.** *Lancet* 2020, **395:**1063-1077.

178   4.   Wang L, Wang Y, Jin S, Wu Z, Chin DP, Koplan JP, Wilson ME: **Emergence and control of infectious diseases in China.** *Lancet* 2008, **372:**1598-1605.

180   5.   Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al: **A new coronavirus associated with human respiratory disease in China.** *Nature* 2020, **579:**265-269.

182   6.   Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al: **A pneumonia outbreak associated with a new coronavirus of probable bat origin.** *Nature* 2020, **579:**270-273.

184   7.   Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H: **The architecture of SARS-CoV-2 transcriptome.** *bioRxiv* 2020.

186   8.   Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom K, Zambon M, Ellis J, Lewis PA, Hiscox JA, Matthews DA: **Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site.** *bioRxiv* 2020.

190   9.   Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L, Druce J, Purcell D, Harty L, Tran T, Roberts J, et al: **Direct RNA sequencing and early evolution of SARS-CoV-2.** *bioRxiv* 2020.

192   10.  Bauch CT, Lloyd-Smith JO, Coffee MP, Galvani AP: **Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future.** *Epidemiology* 2005, **16:**791-801.

194   11.  Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D: **Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis.** *Int J Infect Dis* 2020, **93:**201-204.

197   12.  Breban R, Riou J, Fontanet A: **Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk.** *Lancet* 2013, **382:**694-699.

199   13.  Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS: **Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.** *Science* 2020, **367:**1260-1263.

201

202