Differential methylation as a mediator of COVID-19 susceptibility

Sandra Steyaert¹, Geert Trooskens¹, Joris R. Delanghe², Wim Van Criekinge³

¹ doc.ai Inc. 636 Waverley Street, Palo Alto, CA 94301

² Department of Diagnostic Sciences, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium

³ Biobix, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

Corresponding author: Dr. Sandra Steyaert, PhD (sandra@doc.ai)

Abstract

The COVID-19 outbreak shows a huge variation in prevalence and mortality on geographical level but also within populations¹. The *ACE2* gene, identified as the SARS-CoV2 receptor, has been shown to facilitate the viral invasion and people with higher *ACE2* expression generally are more severely affected^{2,3}. As there is a lot of variability in *ACE2* expression between individuals we hypothesized that differential DNA methylation profiles could be (one of) the confounding factors explaining this variability. Here we show that epigenetic profiling of host tissue, especially in the *ACE2* promoter region and its homologue *ACE1*, may be important risk factors for COVID-19. Our results propose that variable methylation can explain (part of) the differential susceptibility, symptom severity and death rate for COVID-19. Our findings are a promising starting point to further evaluate the potential of *ACE1/2* methylation and other candidates as a predictor for clinical outcome upon SARS-CoV2 infection.

Keywords

SARS-CoV2, DNA methylation, *ACE* genes, COVID-19 susceptibility, epigenetic age

List of Abbreviations

3' UTR	3' untranslated region
5' UTR	5' untranslated region
450K	HumanMethylation450 array
ADT	Androgen deprivation therapy
ACE1	Angiotensin-converting enzyme 1
ACE2	Angiotensin-converting enzyme 2
BMIQ	Beta mixture quantile
CDC	Centers for disease control and prevention
COVID-19	Coronavirus disease 19
D/I	Deletion/Insertion
GEO	Gene expression omnibus
IQR	Interquartile range
M	Methylated intensity
r^2	squared Spearman correlation coefficient
SARS-CoV2	Severe acute respiratory syndrome coronavirus 2
TMPRSS2	Transmembrane protease serine 2
TSS	Transcription start site
U	Unmethylated intensity

Main

Background

In December 2019, a novel human coronavirus now named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) emerged and was responsible for the now global outbreak of potentially severe and fatal atypical pneumonia defined as coronavirus disease 19 (COVID-19)¹.

As of August 15th, 2020, SARS-CoV2 is responsible for more than 21 million detected COVID-19 cases and at least 762k deaths worldwide⁴. The outbreak of the COVID-19 pandemic is especially unnerving because it's hard to predict how the virus will affect any individual person. As symptoms vary a lot between infected persons — if they experience them at all — determining if a person is indeed infected by this novel coronavirus is a surprisingly hard and tricky question to answer without testing⁵.

The Centers for Disease Control and Prevention (CDC) reports that the main symptoms are fever/chills, headache, muscle pain, fatigue, coughing, sore throat and shortness of breath, which can appear somewhere between two and 14 days after exposure to the virus⁶. Confirmed COVID-19 cases have shown a wide range of reported symptoms, going from mild to severe illness. According to the CDC, approximately 80% of infected people present few to mild symptoms, while others have a more severe manifestation of the disease, with extreme cases relying on a ventilator to breathe.

There are a few clear risk factors⁷, including age and general health status but even within these subgroups, there is a huge range in severity. It has been shown that next to age and health status, also socio-economic and environmental factors⁸, sex⁹, and even vitamin D status¹⁰ are associated with the immune response and differential susceptibility. Identifying what factors are responsible for the inter-individual response to the virus is of utmost importance to aid in the identification of population groups at higher risk and to guide more effective strategies and protective measurements.

It should be noted that this difference in individual susceptibility is actually not that unusual for infectious diseases, and thus not unique for COVID-19. The same is true for tuberculosis, malaria and the 'common' flu¹¹. The underlying reason, however, is mostly disease-specific due to the fact that the biological pathways involved in the manifestation of the illness often differ.

Certain genomic variants play a role in the specific immune response to viral infections. The last weeks, efforts were initiated to look into the role of genetics, and if each person's unique genetics may affect

their susceptibility and severity to COVID-19. Especially the interplay between the *ACE* genes and SARS-CoV2 has been of great interest.

The angiotensin-converting enzyme 2 (*ACE2*) gene, which been identified as the SARS-CoV2 receptor, has been shown to facilitate in the viral invasion². While most COVID-19 studies currently focus on *ACE2*, a Belgian research group very recently released data from 33 European countries where they looked at the gene coding for angiotensin-converting enzyme 1 (*ACE1*)¹². Although *ACE2* and *ACE1* share only 42% of amino acid identity, they both act as carboxypeptidases to cleave amino acids from the peptides' carboxyl terminal. The *ACE1* enzyme is characterized by a genetic deletion/insertion (D/I) polymorphism in intron 16, which is associated with alterations in circulating and tissue concentrations of ACE. Their results demonstrate that prevalence and mortality of COVID-19 infection correlate with the D allele frequency of the *ACE1* D/I polymorphism. Another research group recently has launched a similar hypothesis³. Interestingly, the D allele has proven to be associated with a reduced expression of *ACE2*³.

It was already known that expression of *ACE2* is significantly increased amongst people who smoke or suffer from diabetes, hypertension, heart diseases and other conditions¹³, and people with higher *ACE2* expression generally seem also more severely affected by SARS-CoV2¹⁴. Interestingly, this mechanistically implicated *ACE2* gene has been shown to be epigenetically regulated. Taken together with the patient characteristics for severe COVID-19 outcome this substantiates an essential role for epigenetic regulation.

Results

To explore the importance of *ACE2* methylation, publicly available DNA methylation data from 5 research studies (~1000 samples) was used. All data originated from blood from “healthy” people and methylation analysis was performed on Illumina Infinium HumanMethylation450 BeadChip (450K) DNA methylation arrays. As in all Illumina methylation assays, methylation values of each probe are expressed as β values, which range from 0 (no methylation) to 1 (fully methylated). Raw β values were preprocessed, cleaned and subsequently normalized. In next step, the 8 probes falling into the *ACE2* region were fetched and visualized. **Figure 1** displays for each individual 8 probe the corresponding methylation levels per sample. The x- and y-axis show each sample's age and its normalized β value for that position/probe, respectively. Probes are sorted per genomic position in *ACE2* (left to right per row).

Two groups are noticeable: (i) the big cloud of samples marked with circles and a lighter color, and (ii) the smaller cloud of darker triangles. This latter group represents data from a cohort of centenarians. Of special interest are the promoter (flanking) regions. Here, TSS200 covers the region from transcription start site (TSS) to -200 nucleotides upstream; TSS1500 represents the region -200 to -1500 nucleotides upstream of TSS. Probes falling into these regulatory regions are colored in orange, red and purple

(TSS1500/200 + 1stExon/5'UTR) It is apparent that in general upon aging the average methylation levels gradually decline while the spread increases. However, when only looking at the group of centenarians methylation levels appear higher and show a remarkable higher consistency (Table 1).

Figure 1: Chronological age (x-axis) vs normalized β values (y-axis) for 8 methylation probes in ACE2.

Subjects older than 85years (here referred to as 'centenarians') are marked with a triangle and a darker color. Probes are colored as follows: orange=TSS1500; red=TSS200; purple=1stExon and/or 5'UTR; green=Gene Body; blue=3'UTR.

There is a lot of variability in *ACE2* expression between individuals and many variants of *ACE2* are associated with medical conditions such as diabetes, hypertension and cardiovascular disorders¹³. Thus, it could be that such predisposing genetic variants may also contribute to the susceptibility and explain the enormous variability in infection rate, symptom development/severity and death rate. But, while definitely plausible, to date, limited associations have been found between *ACE2* variants and response to SARS-CoV2 infection. As *ACE2* plays a role in the SARS-CoV2 infection pathway and *ACE2* expression seems very variable between individuals, differential DNA methylation of *ACE2* could be one of the confounding factors explaining (part of) this variability.

Table 1: Summary statistics of methylation (β) values per probe.

The first column shows the squared Spearman correlation coefficient (r^2) between methylation and age. The second and third columns display for each probe the median β value as well as the interquartile range (IQR) for the non-centenarian group ($^\circ$) and the centenarian group (*), separately.

	$r^2(\beta - \text{Age})$	Median β	IQR β
cg18458833	-0.41	0.88 $^\circ$ / 0.94*	0.07 $^\circ$ / 0.02*
cg21598868	-0.44	0.76 $^\circ$ / 0.83*	0.16 $^\circ$ / 0.12*
cg18877734	-0.41	0.85 $^\circ$ / 0.90*	0.09 $^\circ$ / 0.04*
cg08559914	-0.40	0.82 $^\circ$ / 0.89*	0.09 $^\circ$ / 0.06*
cg16734967	-0.47	0.62 $^\circ$ / 0.69*	0.17 $^\circ$ / 0.14*
cg05748796	-0.40	0.81 $^\circ$ / 0.88*	0.10 $^\circ$ / 0.04*
cg05039749	0.12	0.10 $^\circ$ / 0.10*	0.05 $^\circ$ / 0.04*
cg23232263	-0.46	0.80 $^\circ$ / 0.92*	0.13 $^\circ$ / 0.04*

Hoffmann and coworkers showed that in addition to ACE2, SARS-CoV2 infection also depends on the host cell factor transmembrane protease serine 2 (TMPRSS2), a cellular protease. While ACE2 is used for initial binding and host cell entry, the spike protein of SARS-CoV2 is primed by TMPRSS2. A process of which the authors demonstrate can be blocked by a clinically proven protease inhibitor¹⁵. The *TMPRSS2* gene is known to be regulated by androgen¹⁶. Interestingly, inhibition of the androgen receptor by for example androgen deprivation therapy (ADT) decreases TMPRSS2 levels¹⁷ – the same effect as higher (promoter) methylation – resulting in less viral priming options and thus less viral entry¹⁸.

DNA methylation is not static. With age, the methylation state of various genes may change¹⁹. It is also known that general health status and lifestyle has an impact on the methylation signatures. These changes are quantifiable and serve as a means to determine one's "epigenetic age" which often differs from the chronological age²⁰. Epigenetic age is a collection or footprint reflecting a combination of a person's genetic make-up but also his/her history of past external experiences/exposures. A compelling example is the case of super-centenarians (people who live > 100 years old and seem to age very healthy). When looking at their epigenetic age, it is significantly less than their chronological age. In these super-centenarians, the underlying reason is probably a combination of both genetic predisposition and lifestyle.

Could differential *ACE2* promoter methylation in host tissue be causative to the differential expression that is seen amongst COVID-19 patients? Could variable DNA signatures of *ACE2* (and *TMPRSS2*) explain (at least part of) the differential susceptibility, symptom development/severity and death rate for COVID-19? Here, we hypothesize that epigenetic profiling, especially DNA methylation signatures, may indeed

permit discovery of population-, age- and gender-related risk factors for COVID-19. The more consistent methylation levels in the centenarian cohort is at least very compelling.

Our findings are a promising starting point to further evaluate the potential of *ACE2* and *TMPRSS2* methylation as a predictor for clinical outcome upon SARS-CoV2 infection. Note that blood DNA represents a mixture of DNA from various leucocyte types. However, due to close contact in the alveoli, methylation patterns found in blood can closely reflect methylation in lung tissue. Other studies focusing on the effect of smoking on DNA methylation indeed found that the effects in blood samples were very similar to the changes in lung tissue^{21, 22}. But to prove any association, follow-up research is needed on respiratory samples from patients. We are now collecting these samples and aim to compare methylation profiles of *ACE2*, *TMPRSS2* and other interesting candidate genes from both COVID-19 negatives and positives (with ranging severity/complications). As the virus continues to spread, more mild cases will arise, and healthcare professionals need to recognize these to accurately portray total numbers of COVID-19 infections. Differentiating mild and moderate from severe disease may also help clinicians in more accurately triaging cases who need medical attention and minimize the risks on the population, health systems, and economy.

Methods

Methylation was measured on Illumina's 450K methylation array for which raw data was downloaded from the Gene Expression Omnibus (GEO)²³ (GSE30870, GSE32149, GSE36064, GSE41169, GSE42861). Details on the individual data sets and characteristics of the study cohort can be found in **Table 2** below. A full description of each dataset can be found in the original reference.

The Illumina 450K BeadChip measures bisulfite-conversion-based, single-CpG resolution DNA methylation levels for over 480K cytosine sites and covers 96% of CpG islands in the human genome. Unlike the previous platform (27K BeadChip) the Illumina 450K BeadChip includes two distinct probe types, Infinium I (n=135,501) and Infinium II (n=350,076). In the Infinium I type, each CpG site is targeted by two 50bp probes: one for detecting the methylated intensity (M) and one for detecting the unmethylated (U) intensity, whereas for Infinium II types, both the M and U intensity of each CpG site are detected by one single probe using different dye colors (green and red). Methylation values per CpG site are indicated by the β -value which ranges from 0 (no methylation) to 1 (fully methylated) and is computed as $\beta = M / (M + U + \alpha)$ where α is 100 generally²⁴.

Raw β -values were preprocessed in R (v3.6.3) with the RnBeads package (v2.4.0)²⁵. Probes not in CpG context were filtered out as well as probes for which the β values were NA or had low variability (standard deviation < 0.005). β -values remaining probes were next normalized using the Beta Mixture Quantile

dilation method (BMIQ)²⁶. In BMIQ, the β -values of type II probes are adjusted into a statistical distribution characteristic of type I probes. After normalization, probes corresponding with *ACE2* were fetched and the β -values of each sample were plotted against the subjects' respective chronological age.

Table 2: Details on the used 450K methylation data sets and characteristics of the study cohort.

ID	DNA origin	Platform	#samples	Median Age (range)	Description
GSE30870	Blood PMBC	450K	40	44 (0,100)	Peripheral blood mononuclear cells from newborns and centenarians.
GSE32149	Blood PMBC	450K	46	15 (3.5,76)	Peripheral blood leukocytes from a DNA methylation study of Crohn's disease and ulcerative colitis. There was no significant evidence that disease status affects DNA methylation.
GSE36064	Blood PMBC	450K	77	3.1 (1,16)	Leukocytes from healthy male children from Children's Hospital Boston consists of peripheral blood leukocyte samples from healthy males
GSE41169	Blood WB	450K	94	29 (18,65)	Whole blood samples from a Dutch population comprised of schizophrenics and healthy control subject. It turned out that schizophrenia status was not related to DNA methylation.
GSE42861	Blood WB	450K	689	54 (17,70)	Whole Blood from rheumatoid arthritis patients. Effect of rheumatoid arthritis on DNA methylation was found to be negligible.

References

1. Zheng, J., *SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat*. Int J Biol Sci, 2020. **16**(10): p. 1678-1685.
2. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin*. Nature, 2020. **579**(7798): p. 270-273.
3. Skarstein Kolberg, E., *ACE2, COVID19 and serum ACE as a possible biomarker to predict severity of disease*. J Clin Virol, 2020. **126**: p. 104350.
4. Dong, E., H. Du, and L. Gardner, *An interactive web-based dashboard to track COVID-19 in real time*. Lancet Infect Dis, 2020. **20**(5): p. 533-534.
5. Xie, J., et al., *Clinical Characteristics of Patients Who Died of Coronavirus Disease 2019 in China*. JAMA Netw Open, 2020. **3**(4): p. e205619.
6. Center of Disease Control and Prevention. *Symptoms of Coronavirus*. [cited 2020 15 June]; Available from: <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html>.

7. Center of Disease Control and Prevention. People who are at higher risk for severe illness. [cited 2020 15 June]; Available from: <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html>.
8. Manrai A, Patel C. COVID-19: Mapping communities at risk. [cited 2020 15 June]; Available from: <https://www.xy.ai/process/covid-19-mapping-communities-at-risk>.
9. Jordan, R.E., P. Adab, and K.K. Cheng, Covid-19: risk factors for severe disease and death. *BMJ*, 2020. **368**: p. m1198.
10. Grant, W.B., et al., Evidence that Vitamin D Supplementation Could Reduce Risk of Influenza and COVID-19 Infections and Deaths. *Nutrients*, 2020. **12**(4).
11. Verhein, K.C., H.L. Vellers, and S.R. Kleeberger, Inter-individual variation in health and disease associated with pulmonary infectious agents. *Mamm Genome*, 2018. **29**(1-2): p. 38-47.
12. Delanghe, J.R., M.M. Speeckaert, and M.L. De Buyzere, The host's angiotensin-converting enzyme polymorphism may explain epidemiological findings in COVID-19 infections. *Clin Chim Acta*, 2020. **505**: p. 192-193.
13. Patel, V.B., et al., Role of the ACE2/Angiotensin 1-7 Axis of the Renin-Angiotensin System in Heart Failure. *Circ Res*, 2016. **118**(8): p. 1313-26.
14. Sawalha, A.H., et al., Epigenetic dysregulation of ACE2 and interferon-regulated genes might suggest increased COVID-19 susceptibility and severity in lupus patients. *Clin Immunol*, 2020. **215**: p. 108410.
15. Hoffmann, M., et al., SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, 2020. **181**(2): p. 271-280 e8.
16. Lucas, J.M., et al., The androgen-regulated protease TMPRSS2 activates a proteolytic cascade involving components of the tumor microenvironment and promotes prostate cancer metastasis. *Cancer Discov*, 2014. **4**(11): p. 1310-25.
17. Ferraldeschi, R., et al., Targeting the androgen receptor pathway in castration-resistant prostate cancer: progresses and prospects. *Oncogene*, 2015. **34**(14): p. 1745-57.
18. Montopoli, M., et al., Androgen-deprivation therapies for prostate cancer and risk of infection by SARS-CoV-2: a population-based study (N = 4532). *Ann Oncol*, 2020.
19. Bell, C.G., et al., DNA methylation aging clocks: challenges and recommendations. *Genome Biol*, 2019. **20**(1): p. 249.
20. Horvath, S., DNA methylation age of human tissues and cell types. *Genome Biol*, 2013. **14**(10): p. R115.
21. Stueve, T.R., et al., Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. *Hum Mol Genet*, 2017. **26**(15): p. 3014-3027.
22. Gao, X., et al., DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics*, 2015. **7**: p. 113.
23. Edgar, R., M. Domrachev, and A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 2002. **30**(1): p. 207-10.
24. Pidsley, R., et al., A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 2013. **14**: p. 293.
25. Muller, F., et al., RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol*, 2019. **20**(1): p. 55.
26. Teschendorff, A.E., et al., A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 2013. **29**(2): p. 189-96.

Availability of data and materials

The datasets analyzed during the current study are available in the Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>) under accession numbers GSE30870, GSE32149, GSE36064, GSE41169 and GSE42861.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SS pre-processed, analyzed and interpreted the methylation data and wrote the manuscript with biological and technical insight from WVC. Additional validation was done by GT. GT and JRD reviewed the analysis and final text. WVC oversaw the work and was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Adriaan Verhelle (Scripps Research, La Jolla, CA, USA) for valuable discussions, comments and help with figure layout.