Retrospective *in silico* HLA predictions from COVID-19 patients reveal alleles associated with disease prognosis

René L. Warren and Inanç Birol

René L. Warren
Genome Sciences Centre, BC Cancer
Vancouver, BC, V5Z 4S6, Canada
rwarren@bcgsc.ca
570 W 7th Ave #100, Vancouver, BC V5Z 4S6
(604) 707-5900 #673235

*To whom correspondence should be addressed

Inanç Birol
Genome Sciences Centre, BC Cancer
Vancouver, BC, V5Z 4S6, Canada
ibirol@bcgsc.ca

ABSTRACT

Background: The Human Leukocyte Antigen (HLA) gene locus plays a fundamental role in human immunity, and it is established that certain HLA alleles are disease determinants.

Methods: By combining the predictive power of multiple *in silico* HLA predictors, we have previously identified prevalent HLA class I and class II alleles, including DPA1*02:02, in two small cohorts at the COVID-19 pandemic onset. Since then, newer and larger patient cohorts with controls and associated demographic and clinical data have been deposited in public repositories. Here, we report on HLA-I and HLA-II alleles, along with their associated risk significance in one such cohort of 126 patients, including COVID-19 positive (n=100) and negative patients (n=26).

Results: We recapitulate an enrichment of DPA1*02:02 in the COVID-19 positive cohort (29%) when compared to the COVID-negative control group (Fisher's exact test [FET] p=0.0174). Having this allele, however, does not appear to put this cohort's patients at an increased risk of hospitalization. Inspection of COVID-19 disease severity outcomes reveal nominally significant risk associations with A*11:01 (FET p=0.0078), C*04:01 (FET p=0.0087) and DQA1*01:02 (FET p=0.0121).

Conclusions: While enrichment of these alleles falls below statistical significance after Bonferroni correction, COVID-19 patients with the latter three alleles tend to fare worse overall. This is especially evident for patients with C*04:01, where disease prognosis measured by mechanical ventilation-free days was statistically significant after multiple hypothesis correction (Bonferroni p = 0.0023), and may hold potential clinical value.

61 Keywords (8): COVID-19, SARS-CoV-2, Human Leukocyte Antigen (HLA), RNA-seq,

62 DPA1*02:02, A*11:01, C*04:01, DQA1*01:02

INTRODUCTION

Modern history has been plagued by deadly outbreaks, from the recurring influenza (e.g. Spanish, Asian, Hong Kong, Avian) and HIV/AIDS viral pandemics, to bacterial and protist infections causing tuberculosis and malaria. Since the early 2000s, we have faced another threat: novel coronavirus infections causing severe respiratory illnesses such as SARS, MERS and today, coronavirus disease 2019 − COVID-19 [1]. The SARS-CoV-2 coronavirus responsible for the COVID-19 respiratory disease is of particular concern; not only does SARS-CoV-2 spread quickly, the symptoms of its infection, when exhibited, are very similar to that of the cold and flu making it difficult to diagnose, trace and contain. Further, infected individuals are affected differently. For instance, older men (≥65 years old) with pre-existing medical conditions, such as diabetes, appear at increased risk of progressing into the more severe phase of the disease, yet SARS-CoV-2 infections affect all other age groups evenly except occasionally in children and adolescents [2]. Most peculiar is that a high proportion of individuals who tested positive for SARS-CoV-2 are asymptomatic – as high as 43% recorded in Iceland, a rate that appears to vary depending on jurisdictions and populations [3-5]. As of now, the disparity in patient response to SARS-CoV-2 infection is still eluding us.

The most efficient way to combat pathogens has been through the use of our own defense mechanism: our acquired immunity. This is done by vaccination campaigns that effectively prime our immune systems at the population level before we even encounter pathogens. But design of effective vaccines must consider interactions

with host immune genes. The Human Leukocyte Antigens (HLA) are a group of such genes encoding surface receptors that bind short peptide epitopes derived from endogenous (class I) or exogenous (class II) antigens, including viral antigens, and they facilitate killer or helper T cells to set off an appropriate immune response. The magnitude of this response varies between patients as populations and individuals have different composition of HLA genes and variable T cell repertoires. As such, HLA induces a bias, which is responsible for documented host susceptibility to disease [6]. Some of the notable associations between HLA and disease are observed in AIDS patients, with certain HLA alleles conferring protection [7]. In other cases, HLA has been implicated with autoimmune diseases and diabetes [8-11]. The exact underlying mechanisms behind these associations are unclear, but there is mounting evidence that bacterial and viral infection may be the trigger for some [10] and that HLA plays a critical role in the viral infection cycle, including viral entry into host cells [12].

Since the beginning of the pandemic, worldwide reports have emerged on HLA associations with COVID-19, including our own [13-19]. Using publicly available metatranscriptomic sequencing data made available at the pandemic onset, we had demonstrated the utility of a high throughput *in silico* method for characterizing the HLA types of COVID-19 patients from bronchoalveolar lavage fluid and blood samples and reported on prevalent alleles, including the DPA1*02:02P - DPB1*05:01P HLA-II haplotype observed in 7 out of the 8 of patients from two small cohorts. Here, using public RNA-seq sequencing data from a larger COVID-19 patient cohort with clinical outcomes and demographics data [20], we report on HLA alleles with potential diagnostic (DPA1*02:02) and prognostic (A*11:01, C*04:01 and DQA1*01:02) value in

126 hospitalized patients with (n=100) and without (n=26) COVID-19 and present our findings in light of available demographic characteristics using hospitalization and disease severity metrics.

METHODS

We downloaded Illumina NOVASEQ-6000 paired-end (50 bp) RNA-Seq reads from libraries prepared from the blood samples of 126 hospitalized patients, with (n=100) or without COVID-19 (n=26) (ENA project: PRJNA660067, accessions: SRX9033799-SRX9033924). This data is part of a large-scale multi omics study from the Department of Molecular and Cellular Physiology, Albany Medical College, Albany, NY, USA, with aims to analyze COVID-19 Severity and clinical as well as demographics data was made available by the study authors [20] (GEO accession GSE157103). On each patient RNA-Seq dataset, we ran HLA prediction software OptiType [21] (v1.3.4), seq2HLA [22] (v2.3), and HLAminer [23] (v1.4 targeted assembly mode with defaults) as described [19]. We tallied HLA class I (HLA-I) and class II (HLA-II, supported by Seq2HLA and HLAminer only) allele predictions and for each patient we report the most likely HLA allele (4-digit resolution), indicating HLA predictor tool support (**Additional file 1, tables S1 and S2**).

Looking at class I and II alleles predicted in 10% or more of COVID-19 positive patients (class I, n=17; class II, n=11) we calculated Fisher's Exact Test (FET), first testing for enrichment in COVID-19 positive vs. negative patients (R function fisher.test, alternative = "greater"). For those same alleles (found in ≥10% patients) and inspecting only the COVID-19 positive cohort, we tested for the probability of patient hospitalization, as measured by the Intensive Care Unit (ICU) metric reported by the original study authors, using FET. We looked further into the risk of hospitalization in COVID-19 patients with vs. without these alleles using the Kaplan-Meier (KM) estimator (R library survival), plotting the probability of remission using the "hospital-free days

post 45 day followup (days)" (HFD-45) metric reported by the study author as a proxy for disease severity, with lower HFD-45 numbers indicating worse outcomes. Similarly, we ran the KM estimator using another metric of disease severity, "ventilator-free days", which captures the most severe cases with COVID-19 patients suffering respiratory deterioration and requiring mechanical ventilation. On each set we calculated the log-rank p-value (R library survminer) and corrected for multiple hypothesis testing (Bonferroni correction) using the number and patient abundance rank of class I (n=17) or class II (n=11) HLA alleles observed in 10% or more of COVID-19 patients. We also inspected the combined influence of HLA alleles and patient demographics data (age, sex, ethnicity) on the hospitalization (ICU negative vs. positive) outcomes of COVID-19 patients, using odds-ratio calculations (R function fisher.test, and applying Haldane correction [24] on zero values, when necessary).

RESULTS

We collated the HLA class I and class II predictions of three *in silico* HLA predictors derived from the RNA-seq samples of a recent [20] COVID-19 positive patient cohort (n=100) with control patients (n=26) who tested negative for COVID-19 (**Additional file 1, Tables S1 and S2**). Due to the limiting short read length (paired 50 bp) we chose to first report on OptiType [21] and seq2HLA [22] class I and class II predictions, and count the additional allele support from seq2HLA and HLAminer [23]. In all, we identify 17 and 11 HLA class I and class II alleles predicted in 10% or more of COVID-19 patients, respectively (**Tables 1 and 2**). There were many more alleles predicted (133 class I and 101 class II in the COVID-19 positive cohort), but too few patients are represented at lower cut-offs to compute meaningful statistics. First, we looked at the statistical enrichment (Fisher's Exact Test - FET) of each allele in the COVID-19 positive set, compared to the COVID-19 negative control group. We find HLA-I A*30:02 and HLA-II DPA1*02:02 allele enrichments nominally significant (FET p = 0.0417 and p = 0.0174) at the 5% level (**Tables 1 and 2**). However, when Bonferroni correction is applied for the number of HLA class I allele tests or when the abundance rank is factored in for A*30:02, the test is not significant (n=17, Bonferroni p = 0.7080; n=10, Bonferroni p = 0.4165, respectively). For HLA-II DPA1*02:02, Bonferroni correction finds the test insignificant at the α=0.05 level for the number of hypothesis, but significant for the allele abundance rank (n=11, Bonferroni p = 0.1916; n=2, Bonferroni p = 0.0348).

COVID-19 positive patients could be further stratified into those who were hospitalized and admitted to the Intensive Care Unit (**Tables I and 2**, ICU+), and those who were not (**Tables 1 and 2**, ICU-). When computing FET statistics, we find HLA-I

A*11:01 and C*04:01 and HLA-II DQA1*01:02 significant at the α=0.05 level (**Tables 1 and 2**; p = 0.0078, p = 0.0087 and 0.0121, respectively) but none remain significant after Bonferroni correction. The Overmyer study authors [20] reported important disease severity metrics (HFD-45 and days without needing mechanical ventilation), which we used to assess the remission probability of COVID-19 patients having a specific allele using Kaplan-Meier estimation. We find patients of the Overmyer cohort with either A*11:01 (**Figure1a**), C*04:01 (**Figure 1b**) or DQA1*01:02 (**Figure 1c**) to be at a significant increased risk of hospitalization (log-rank p = 0.0099, p = 0.0082 and p = 0.0097). When applying multiple test corrections to account for allele abundance rank, only C*04:01 (n=5, Bonferroni p = 0.0410) and DQA1*01:02 (n=2, Bonferroni p = 0.0194) remained significant at the α=0.05 level. When looking at patients needing mechanical ventilators, a severe outcome in COVID-19 disease progression, we only find patients with C*04:01 to be at a statistically significant increased risk (**Figure 1d**, log-rank p = 0.0019). Multiple hypothesis test correction retains the statistical significance of this allele when factoring both the number of HLA-I alleles tested and C*04:01 abundance rank (n=17, Bonferroni p = 0.0023; n=5, Bonferroni p = 0.0095).

Looking at the influence of the aforementioned alleles in combination with simple demographics (sex, age and ethnicity), we find that of the Overmyer cohort patients with the DPA1*02:02 allele, those with a white ethnic background and females appear at an increased risk of testing positive for COVID-19 (**Figure 2**; odds ratio [OR] = 6.33 [5.33–7.34], FET p = 0.0491 and OR = 7.33 [6.18-8.48], FET p = 0.0326, respectively). The association with gender is also observed in alleles A*11:01, C*04:01 and DQA1:01:02, putting female COVID-19 patients of this cohort at an increased risk of hospitalization

for the class I alleles (**Figure 2**; OR = 12.09 [10.41-13.76], FET p = 0.0105 for both) and male COVID-19 patients at increased risk for DQA1*02:02 (**Figure 2**; OR = 2.74 [2.69-2.79], FET p = 0.0481). For patients with the latter HLA-II allele, minorities and younger individuals (<65 years old) are also more at risk of hospitalization (**Figure 2**; OR = 4.08 [2.85-5.32], FET p = 0.0222 and OR = 3.62 [2.42-4.83], FET p = 0.0240, respectively). In this cohort, we also find patients with A*11:01 in the younger age group (<65 years old) at increased risk of hospitalization (**Figure 2**; OR = 9.54 [8.14-10.94], FET p = 0.0184) whereas for those with C*04:01, it appears a white ethnic background and a more advanced age (≥65 years old) may be predisposing to ICU hospitalization (**Figure 2**; OR = 14.25 [12.24-16.26], FET p = 0.0053 and OR = 9.66 [8.27-11.05], FET p = 0.0188, respectively).

DISCUSSION

We have previously identified the DPA1*02:02 class II allele as being prevalent in two other and independent cohorts, with patients of undisclosed ethnic background, but hospitalized in Wuhan, China [19]. Of populations with reported allele frequencies and an appreciable sampling size (≥100 individuals), only Hong Kong Chinese and Japanese have DPA1*02:02 allele frequencies (55.8% and 43.5%, respectively; [25,26]) above its observed frequency (29.0%) in the COVID-19 positive cohort analyzed herein. The frequency of this allele in other qualifying populations tends to be generally lower, including in South African (Worcester, 15.6%), Norwegians (14.0%), Mexico Chiapas Lacandon Mayans (6.7%), United Kingdom Europeans (4.3%) and Spain Navarre Basques (2.2%). We note that the ethnic background of the Overmyer *et al* [20] cohort is heterogeneous, and white individuals (of unknown ancestry) represent 45.1% and (11/29) 37.9% of the COVID-19 positive cohort and its DPA1*02:02 subset, respectively. In contrast, Asians represent only a minority of the cohort (1.9%) and DPA1*02:02 subset (1/29~3.4%). It is important to note that, in the Overmyer cohort, DPA1*02:02 is not statistically associated with increased risk of hospitalization. The significant enrichment of this allele in the COVID-19 positive vs. negative cohorts (FET p= 0.0174) across all individuals, but also when looking only at females (FET p= 0.0326) or white individuals (FET p= 0.0491), and not any other demographics, may prove an important disease marker, which would need to be validated with additional datasets and in independent studies.

There are reports of disease associations with DPA1*02:02, DQA1*01:02, C*04:01 and A*11:01, but they are few. Of note, the association of all aforementioned alleles with narcolepsy [10,11] and a known trigger for this auto-immune disease includes upper-airway infections and influenza vaccinations [27-32]. Susceptibility to narcolepsy may in fact be an indirect effect of HLA class I and the HLA class II DP isotype in response to viral and bacterial infections, including from influenza and *streptococcus* [10,27,33,34]. It has since been reported that HLA-A*11 may be a susceptibility allele to influenza A(H1N1)pdm09 infection in some populations [35] while another report implicates HLA-I allele C*04:01 with high HIV viral loads [36]. Further, it was recently demonstrated that MHC class II DR, DQ and DP isotypes play a role in mediating the cross-species entry of bat influenza viruses *in vitro* in human/animal cell lines and in mice where engineered MHC-II deficiency made them resistant to upper-respiratory tract infections [12]. It is therefore not a stretch to envision an involvement from these HLA class II isotypes in controlling the cellular entry of a broader range of viral agents *in vivo*.

In a recent study examining HLA susceptibility based on SARS-CoV-2 derived peptide (epitope) binding strengths [37], the HLA-I allele A*11:01 was *in silico* predicted to bind a large number of SARS-CoV-2 derived peptides (n=750) with varying affinity [IC50 range 4.95 – 498.19, median = 149.62, mean = 182.28], and has been experimentally validated to bind SARS-CoV-2 peptide GLMWLSYFV (Tables S4 and S7 in Nguyen *et al.* [37]). In contrast, C*04:01 was only predicted *in silico* to bind six SARS-CoV-2 peptides and at higher IC50 ranges [167.65 – 469.30, median = 291.06, mean =

299.01] (Table S7 in Nguyen *et al.* [37]) suggesting a more limited ability to present epitopes to T cells and mount an appropriate immune response.

There have been a number of reports published on HLA alleles − COVID-19 associations this past year, and on cohorts from many jurisdictions including China [18], Italy [13,14,16] and the UK [17]. Wang *et al*. [18] compared the HLA allele frequencies between a cohort of 82 Chinese individuals and a control population of bone marrow donors previously studied by the same group. Novelli and co-workers [14] HLA typed a cohort of 99 Italian COVID-19 patients, and associated the observed allele frequencies with the HLA types in a reference group of 1,017 Italian individuals also previously studied by the same group. Correale *et al*. [13] and Pisanti *et al.* [16] followed a different strategy; these two independent studies leveraged population scale genomics data retrieved from the Italian Bone-Marrow Donors Registry and the National Civil Protection Department. They correlated background HLA allele frequency data with mortality and morbidity rates across Italy to reach at starkly different conclusions on which HLA alleles may play a role in disease etiology and progression. Disagreement between these two studies (also distinct from the results of the other Italian study by Novelli *et al*.) highlight the importance of large cohorts with matched samples to infer the patient HLA alleles with better statistical significance. Poulton *et al*. [17] characterized the HLA types of 80 COVID-19 patients in the UK on waiting lists for transplantation, and compared observed allele frequencies in comparison to a cohort of 10,000 deceased organ donors and a separate cohort of 308 SARS-CoV-2-negative individuals also on waiting lists for transplantation, the latter representing a matched demographics for the COVID-19 patients in their cohort. Interestingly, this is the only

study that had any overlap between the alleles they flagged as being statistically significant and the lists published by other studies cited above. Not surprisingly, the alleles they listed do not intersect with the alleles identified and presented herein and the three alleles we published earlier on a very small group of only eight patients. It is nonetheless intriguing to find little to no HLA allele overlap between these reports, including with those associated with the 2003 SARS outbreak, a related respiratory disease caused by coronavirus [15,38,39]. This could be explained, at least partially, by geographical differences and varying population allele frequencies in those cohorts, relatively small cohort sizes (<100 patients), differences in experimentation setup and/or other factors, including comorbidity status, that may be acting independently of HLA.

CONCLUSIONS

Here, we predict HLA-I and HLA-II alleles from publicly available COVID-19 patient blood RNA-seq samples and identified several putative biomarkers. In a previous study, we had observed one of these biomarkers (DPA1*02:02), and we postulate that patients with the allele may have an increased susceptibility for COVID-19. Further, other alleles (A*11:01, C*04:01, DQA1*01:02) may be prognostic indicators of poor outcome. However, although it is well established that patient HLA profiles play a significant role in the onset and progression of infectious diseases in general, we caution against drawing overreaching conclusions from regional, and often limited, observations. We note that recently published studies associating HLA alleles and COVID-19, by and large, disagree in their findings. We expect future studies with larger cohort sizes will help bring a clearer picture of the role of patient HLA profiles, if any, in COVID-19 susceptibility and disease outcomes.

LIST OF ABBREVIATIONS

FET: Fisher's Exact Test; HFD-45: hospital-free days post 45 day followup (days); HLA: human leukocyte antigen; ICU: Intensive Care Unit; KM: Kaplan-Meier

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The RNA-seq datasets analysed during the current study are available in the ENA repository https://www.ebi.ac.uk/ena/browser/view/PRJNA660067 accessions: SRX9033799- SRX9033924. The associated clinical data are available in the GEO repository https://www.ncbi.nlm.nih.gov/geo accession: GSE157103

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by Genome BC and Genome Canada [281ANV]; and the National Institutes of Health [2R01HG007182-04A1]. The content of this paper is solely the responsibility of the authors, and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.

Authors' contributions

RLW designed the study, analyzed the data and wrote the manuscript. IB participated in the development of the study and co-wrote the manuscript. All authors have read and approved the manuscript.

Acknowledgements

Not applicable

REFERENCES

1. Petersen E, Koopmans M, Go U, Hamer DH, Petrosillo N, Castelli F, et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect Dis. 2020;20:e238-e244.

2. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020;382:1708-20.

3. Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, et al. Spread of SARS-CoV-2 in the Icelandic Population. N Engl J Med. 2020;382:2302-15.

4. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Euro Surveill. 2020;25:2000180.

5. Nishiura H, Kobayashi T, Miyama T, Suzuki A, Jung SM, Hayashi K, et al. Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19). Int J Infect Dis. 2020;94:154-5.

6. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018;18:325-39.

7. Goulder PJ, Walker BD. HIV and HLA class I: an evolving relationship. Immunity. 2012;37:426-40.

8. Chu X, Yang M, Song ZJ, Dong Y, Li C, Shen M, et al. Fine mapping MHC associations in Graves' disease and its clinical subtypes in Han Chinese. J Med Genet. 2018;55:685-92.

9. Noble JA, Valdes AM, Bugawan TL, Apple RJ, Thomson G, Erlich HA. The HLA class I A locus affects susceptibility to type 1 diabetes. Hum Immunol. 2002;63:657-64.

10. Ollila HM, Ravel JM, Han F, Faraco J, Lin L, Zheng X, et al. HLA-DPB1 and HLA class I confer risk of and protection from narcolepsy. Am J Hum Genet. 2015;96:136-46.

11. Tafti M, Lammers GJ, Dauvilliers Y, Overeem S, Mayer G, Nowak J, et al. Narcolepsy-Associated HLA Class I Alleles Implicate Cell-Mediated Cytotoxicity. Sleep. 2016;39:581-7.

12. Karakus U, Thamamongood T, Ciminski K, Ran W, Günther SC, Pohl MO, et al. MHC class II proteins mediate cross-species entry of bat influenza viruses. Nature. 2019;567:109-12.

13. Correale P, Mutti L, Pentimalli F, Baglio G, Saladino RE, Sileri P, Giordano A. HLA-B*44 and C*01 Prevalence Correlates with Covid19 Spreading across Italy. Int J Mol Sci. 2020;21:5205.

14. Novelli A, Andreani M, Biancolella M, Liberatoscioli L, Passarelli C, Colona VL, et al. HLA allele frequencies and susceptibility to COVID-19 in a group of 99 Italian patients. HLA. 2020; doi:10.1111/tan.14047.

15. Ovsyannikova IG, Haralambieva IH, Crooke SN, Poland GA, Kennedy RB. The role of host genetics in the immune response to SARS-CoV-2 and COVID-19 susceptibility and severity. Immunol Rev. 2020;296:205-19.

16. Pisanti S, Deelen J, Gallina AM, Caputo M, Citro M, Abate M, et al. Correlation of the two most frequent HLA haplotypes in the Italian population to the differential regional incidence of Covid-19. J Transl Med. 2020;18:352.

17. Poulton K, Wright P, Hughes P, Savic S, Welberry Smith M, Guiver M, et al. A role for human leucocyte antigens in the susceptibility to SARS-Cov-2 infection observed in transplant patients. Int J Immunogenet. 2020;47:324-8.

18. Wang W, Zhang W, Zhang J, He J, Zhu F. Distribution of HLA allele frequencies in 82 Chinese individuals with coronavirus disease-2019 (COVID-19). HLA. 2020;96:194-6.

19. Warren RL, Birol I. HLA predictions from the bronchoalveolar lavage fluid and blood samples of eight COVID-19 patients at the pandemic onset. Bioinformatics. 2020; doi:10.1093/bioinformatics/btaa756.

20. Overmyer KA, Shishkova E, Miller IJ, Balnis J, Bernstein MN, Peters-Clarke TM, et al. Large-Scale Multi-omic Analysis of COVID-19 Severity. Cell Syst. 2020; doi:10.1016/j.cels.2020.10.003.

21. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30:3310-6.

22. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J, Boisguérin V, et al. HLA typing from RNA-Seq sequence reads. Genome Med. 2012;4:102.

23. Warren RL, Choe G, Freeman DJ, Castellarin M, Munro S, Moore R, Holt RA. Derivation of HLA types from shotgun sequence datasets. Genome Med. 2012;4:95.

24. Valenzuela, C. Dos soluciones para la estimación de "odds ratios" con ceros [2 solutions for estimating odds ratios with zeros]. Rev Med Chil. 1993;121:1441-4.

25. Gonzalez-Galarza FF, McCabe A, Santos EJ, Jones J, Takeshita LY, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. Nucleic Acid Research. 2020;48:D783-8.

26. Allele frequency net database. http://allelefrequencies.net/

27. Han F, Lin L, Warby SC, Faraco J, Li J, Dong SX, et al. Narcolepsy onset is seasonal and increased following the 2009 H1N1 pandemic in China. Ann Neurol. 2011;70:410-7.

28. Han F, Faraco J, Dong XS, Ollila HM, Lin L, Li J, et al. Genome wide analysis of narcolepsy in China implicates novel immune loci and reveals changes in association prior to versus after the 2009 H1N1 influenza pandemic. PLoS Genet. 2013;9:e1003880.

29. Heier MS, Gautvik KM, Wannag E, Bronder KH, Midtlyng E, Kamaleri Y, Storsaeter J. Incidence of narcolepsy in Norwegian children and adolescents after vaccination against H1N1 influenza A. Sleep Med. 2013;14:867-71.

30. Moss AJ, Gaughran FP, Karasu A, Gilbert AS, Mann AJ, Gelder CM, et al. Correlation between human leukocyte antigen class II alleles and HAI titers detected post-influenza vaccination. PLoS One. 2013;8:e71376.

31. Montplaisir J, Petit D, Quinn MJ, Ouakki M, Deceuninck G, Desautels A, et al. Risk of narcolepsy associated with inactivated adjuvanted (AS03) A/H1N1 (2009) pandemic influenza vaccine in Quebec. PLoS One. 2014;9:e108489.

32. Wijnans L, Lecomte C, de Vries C, Weibel D, Sammon C, Hviid A, et al. The incidence of narcolepsy in Europe: before, during, and after the influenza A(H1N1)pdm09 pandemic and vaccination campaigns. Vaccine. 2013;31:1246-54.

33. Aran A, Lin L, Nevsimalova S, Plazzi G, Hong SC, Weiner K, et al. Elevated anti-streptococcal antibodies in patients with recent narcolepsy onset. Sleep. 2009;32:979-83.

34. Mahlios J, De la Herrán-Arita AK, Mignot E. The autoimmune basis of narcolepsy. Curr Opin Neurobiol. 2013;23:767-73.

35. Dutta M, Dutta P, Medhi S, Borkakoty B, Biswas D. Polymorphism of HLA class I and class II alleles in influenza A(H1N1)pdm09 virus infected population of Assam, Northeast India. J Med Virol. 2018;90:854-60.

36. Olvera A, Pérez-Álvarez S, Ibarrondo J, Ganoza C, Lama JR, Lucchetti A, et al. The HLA-C*04: 01/KIR2DS4 gene combination and human leukocyte antigen alleles with high population frequency drive rate of HIV disease progression. AIDS. 2015;29:507-17.

37. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, Thompson RF. Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. J Virol. 2020;94:e00510-20.

38. Lin M, Tseng HK, Trejaut JA, Lee HL, Loo JH, Chu CC, et al. Association of HLA class I with severe acute respiratory syndrome coronavirus infection. BMC Med Genet. 2003;4:9.

39. Sanchez-Mazas A. HLA studies in the context of coronavirus outbreaks. Swiss Med Wkly. 2020;150:w20248.

480 **TABLES**

481

482 **Table 1**. HLA-I alleles identified in 10% or more COVID-19 positive patients and
483 statistical tests of enrichment in the Overmyer *et al.* [20] COVID-19 positive (vs.
484 negative) cohort and association with hospitalization. Red font indicates significant
485 associations (Fisher's Exact Test) not corrected for multiple hypothesis tests.

HLA-I	COVID+ patients	ICU- patients	ICU+ patients	COVID- patients	H1: Enriched in COVID+ Fisher's Exact Test p-value	H1: Increased risk of hospitalization Fisher's Exact Test p-value
*A*02:01*	30	15	15	12	0.9614	0.5862
*A*24:02*	23	12	11	6	0.6161	0.5000
*C*07:02*	22	9	13	7	0.7890	0.8865
*C*07:01*	18	6	12	7	0.8991	0.9668
*C*04:01*	18	4	14	3	0.3231	0.0087
*C*06:02*	16	8	8	6	0.8708	0.6071
*B*51:01*	16	6	10	2	0.2292	0.9143
*C*03:04*	14	4	10	3	0.5171	0.9796
*C*15:02*	14	6	8	1	0.1363	0.8060
*A*01:01*	13	5	8	7	0.9742	0.8832
*A*03:01*	13	7	6	5	0.8680	0.5000
*A*30:02*	13	7	6	0	0.0417	0.5000
*B*07:02*	12	5	7	5	0.8969	0.8217
*A*30:01*	12	8	4	1	0.2016	0.1783
*B*44:02*	11	4	7	4	0.8320	0.9001
*A*68:01*	11	5	6	0	0.0697	0.7377
*A*11:01*	10	1	9	2	0.5320	0.0078

486

487

488

Table 2. HLA-II alleles identified in 10% or more COVID-19 positive patients and statistical tests of enrichment in the Overmyer *et al.* [20] COVID-19 positive (vs. negative) cohort and association with hospitalization. Red font indicates significant associations (Fisher's Exact Test) not corrected for multiple hypothesis tests.

HLA-II	COVID+ patients	ICU- patients	ICU+ patients	COVID- patients	H1: Enriched in COVID+ Fisher's Exact Test p-value	H1: Increased risk of hospitalization Fisher's Exact Test p-value
*DPA1*01:03*	46	21	25	17	0.9812	0.8421
*DQA1*01:02*	40	26	14	10	0.5363	0.0121
*DPB1*04:01*	22	9	13	10	0.9726	0.8865
*DPA1*02:02*	29	16	13	2	0.0174	0.3299
*DPA1*02:01*	22	11	11	6	0.6578	0.5952
*DQA1*05:02*	19	9	10	3	0.2824	0.6945
*DPB1*01:01*	19	10	9	3	0.2824	0.5000
*DQB1*06:11*	11	6	5	3	0.6811	0.5000
*DRB1*13:01*	10	7	3	4	0.8689	0.1589
*DQB1*06:02*	10	6	4	3	0.7348	0.3703
*DRB1*15:01*	10	4	6	2	0.5320	0.8411

FIGURES

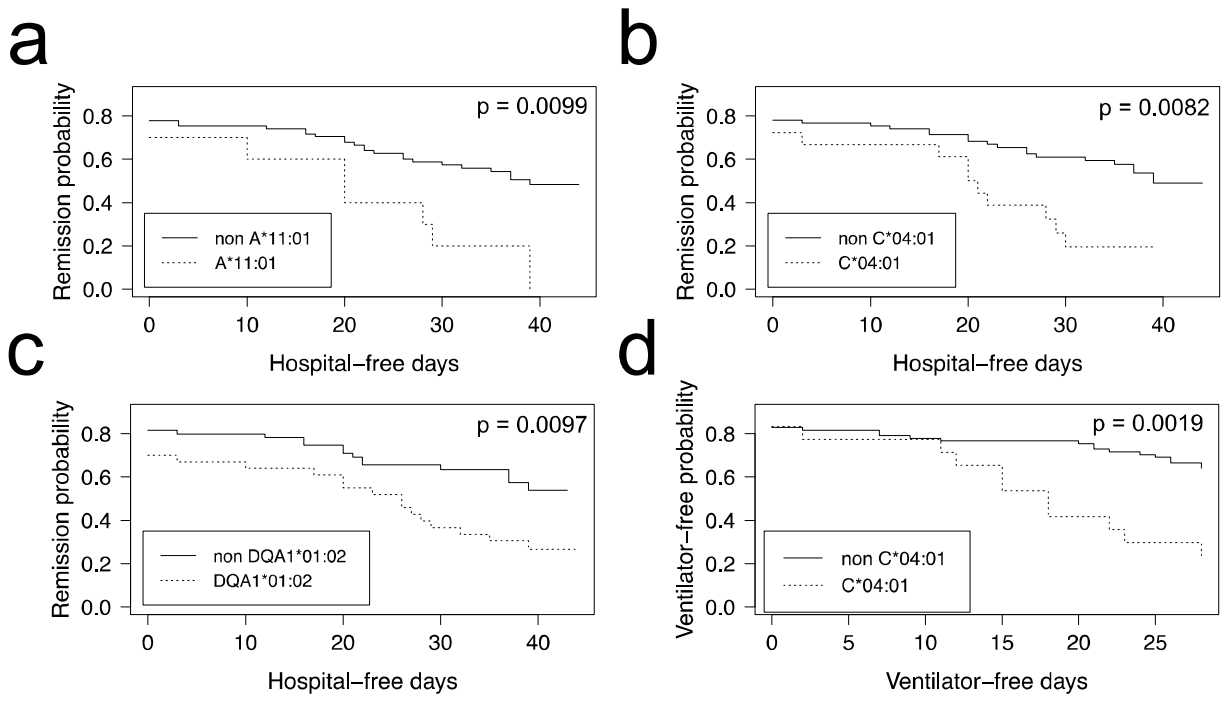

Figure 1. HLA alleles associated with higher risk of hospitalization in a COVID-19 positive patient cohort. COVID-19 positive patients were split into two groups per allele tested, depending on whether they were predicted to have the HLA allele under scrutiny or not. We ran the Kaplan-Meier estimator (R package survival) using the Overmyer *et al.* cohort [20] HFD-45 metric for estimating the remission probability of patients without or with alleles (a) A*11:01, (b) C*04:01 or (c) DQA1*01:02 and mechanical ventilator-free days to estimate the statistical significance of the more severe disease outcome observed in COVID-19 patients with (d) C*04:01. Log-rank p values were calculated for each (R package survminer) and are indicated on the plots.

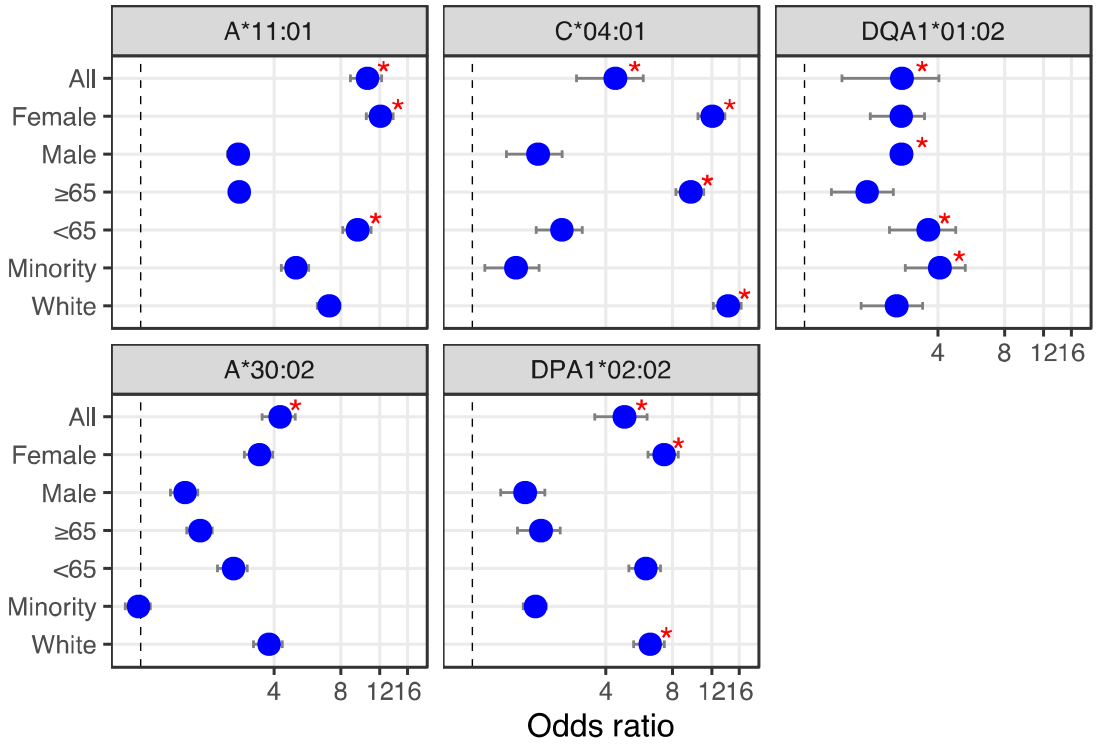

Figure 2. HLA alleles – demographics combinations with diagnostic (bottom) or prognostic (top) potential in a COVID-19 cohort. We calculated the odds ratio (OR) for each HLA-I and HLA-II alleles observed in 10% or more of patients, and plotted OR and the influence of demographics for HLA alleles showing significant associations (from **Tables 1 and 2**). First, looking at the influence of demographic characteristics such as sex (male/female), age (65 years old or above/less than 65 years old) and ethnicity (minority/white ethnic background) on the susceptibility of patients with these alleles to test positive for COVID-19 (lower two panels), and on the risk associated with ICU hospitalization (upper three panels). Red asterisks indicate significant demographic characteristics (Fisher's Exact Test) not corrected for multiple hypothesis tests.