Why do we die? Uncovering the genetic connections causing COVID-19 disease severity

Keyana Cardoza
University of Tampa

Magen Hoch
University of Tampa

Frederic B. Montz
University of Tampa

Bridgette F. Froeschke
University of Tampa  https://orcid.org/0000-0001-9504-7298

Kimberly P. Dobrinski  (kdobrinski@ut.edu)
University of Tampa  https://orcid.org/0000-0002-2626-1204

Research Article

Keywords: COVID-19, Single Nucleotide polymorphisms, eQTL, risk, disease severity

DOI: https://doi.org/10.21203/rs.3.rs-33174/v1

License: ☒  This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

SARS-CoV-2 attacks randomly causing deaths among certain populations. Additionally, this disease seems to kill males preferentially. This study investigates why by examining single nucleotide polymorphisms (SNPs) correlated with increased protein expression associated with viral infection of cells leading to disease proliferation. ACE2, the assumed host cell receptor for the virus is believed to be assisted by co-receptors ENPEP and ANPEP. TMPRSS2 cleaves the S viral protein into two sub-units allowing viral binding to the ACE2 receptor. ACE2 found on the X-chromosome has SNPs increasing ACE2 expression found at frequencies greater than 50% in all male populations analyzed which could account for the increase in male deaths. Females would undergo X-inactivation for the SNPs and have protection from the increased ACE2 expression in all their cells. ACE2, ENPEP, and TMPRSS2 were also found to have population specific SNP patterns which could account for the increased prevalence of disease among certain populations.

Introduction

Severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) causes COVID-19, a disease that has spread into a worldwide pandemic. As of April 19th, 2020, the John Hopkins coronavirus resource center reports 2,394,291 cases of SARS-CoV-2 [1]. The World Health Organization (WHO), weekly surveillance report for Europe had 50% male cases and a 60% male mortality rate [2]. The United States has 755,533 cases, by far the most of any other nation.10 The United States reports 38,664 deaths, a 5.3% case-fatality [1].

Coronaviruses utilize the S protein to bind to host cells facilitating entry and infection [3-6]. The S protein is cleaved into two subunits, S1 and S2 exposing the receptor binder domain (RBD) which is the mechanism for viral cell entry [4-6]. SARS- CoV-2 receptor binding domain (RBD) on S1 utilizes the angiotensin converting enzyme-2 (ACE2) receptor for attachment and entry to host cells. This process is mediated by the S2 subunit. Modifications in RBD show SARS-CoV-2 has 10x more affinity for ACE2 compared to SARS-COVID1 [7].

ACE2 is a negative regulator of the renin-angiotensin system [7]. The ACE2 receptor has been reported to be expressed in the lungs, heart, kidney, ileum, brain, liver, esophagus, colon, nasal goblet and nasal ciliated cells [7-9]. Higher levels of ACE2 expression have found in tissues associated with increased SARS-CoV-2 infection [8,9]. Lung cells of males have 5 ACE2 receptors compared to 2-4 in females and show a higher infection, severity, and mortality rates than females [7,10]. ACE2 in the lungs is suggested to be expressed in type II alveolar (AT2), type I apical epithelial, airway epithelial, fibroblasts, macrophages, esophagus, colon, ileum, nasal goblet and ciliated cells [7-10]. ACE2 receptors in the lungs increase with age [11].

Reports of lower levels of ACE2 in AT2 cells suggest SARS-CoV-2 utilizes co- receptors to facilitate entry [5,8]. Reports suggest glutamyl aminopeptidase (ENPEP), alanine aminopeptidase (ANPEP), and DPP4 may serve as co-receptors [8]. ENPEP is involved with the RAS system, associated with hypertension, and has higher expression in the kidneys [12-15]. Cardiac and kidney complications have been reported with SARS-CoV-2 patients. ENPEP receptors are suggested to play an important role in the RAS system in the brain
Neurological and central nervous system complications in patients have been reported [18]. ANPEP may serve as a binding co-receptor for coronaviruses and is expressed in lung, kidney, and intestine [8,19]. SARS-CoV-2 patients have reported acute kidney and cardiac issues [20]. Studies suggest transmembrane protease serine 2 (TMPRSS2) cleaves the S protein prior to binding to ACE2 and both proteins were shown to be co-expressed in lungs [21]. TMPRSS2 has been suggested to have higher expression than ACE2 [9]. TMPRSS2 and ACE2 were shown to be co-expressed in the esophagus, colon, ileum, nasal goblet and ciliated cells [9]. Co-receptor binding of SARS-CoV-2 may explain the random sudden onset of cardiac, kidney, and neurological symptoms in non-comorbid patients.

The ACE2 receptor is found on the X-Chromosome [22]. X-Linked inactivation (XCI) is the process in female cells of randomly inactivating one X chromosome for proper gene dosage between X-linked genes [23,24]. The maternal or paternal choice of inactivation is random between inherited X chromosomes and the processed genes are passed on to offspring cells [23]. Heterozygous females carrying one wild type and one mutant allele will always have some cellular populations expressing the wild type allele and some cellular populations expressing the mutant allele [23,24]. This may be mixed within tissues. Reports of XCI involving ACE2 suggest variable heterozygous sex bias [22].

Single nucleotide polymorphisms (SNPs) and structural variants (SVs) are a common form of genetic variation. SNPs found near a gene regulatory region may affect gene expression. SVs and SNPs have been associated in genetic disorders such as autism, schizophrenia, and male infertility disorders [25-28]. Differential expression linked to SNPs across divergent world populations may reflect the prevalence of ACE2 receptors, (the SARS-CoV-2 mechanism of entry) within tissues. Males with ACE2 polymorphisms have risk factors for essential hypertension. Increased ACE2 expression, and SVs of ACE2 were found to increase or decrease binding affinity to the SARS-CoV-2 S1 protein [29-31]. Changes in allelic expression may contribute to the higher SARS-CoV-2 infection and mortality rate in males.

COVID-19 is a global pandemic that has caused millions to fall ill and thousands to die. People around the globe want to know why does this disease randomly affect some populations and not others? Why do males seem to be dying at a higher frequency than females? This study aims to address these two puzzling questions and elucidate possible sex and population risk profiles using SNPs found to increase expression of ACE2, TMPRSS2, and ENPEP across tissues. Sex and population differences between frequency of these SNPs may lead to increased expression of these important proteins leading to heightened infection and disease severity.

Results

Death and infection frequencies

Males are dying at a higher frequency than females on a global scale. It is not expected that males would succumb to COVID-19 at a higher frequency than females; however, when compared to a 50% expected death frequency and a 50% expected infection frequency, it was found that New York reported a significant p-value and an expected probability greater than 0.50 for both male infection and male death by COVID-19.
(Table 1a). Three states were found to have a significant p-value and an expected probability greater than 0.50 for male infections by COVID-19: New Jersey, Utah, and South Dakota (Table 1b). Twelve states were found to have a significant p-value and an expected probability greater than 0.50 for male deaths by COVID-19 when compared to 50% expected probability: Arizona, Colorado, Florida, Illinois, Indiana, Michigan, North Carolina, Ohio, Pennsylvania, Virginia, Washington, and Wisconsin (Table 1c).

Table 1a- Statewide COVID-19 binomial testing results for both male infections and male deaths that reported a significant p value and an expected probability greater than 0.50

| State      | Chi Square Statistic for Male Infection | P Value for Male Infection | Chi Square Statistic for Male Death | P Value for Male Death | n, the total number of deaths |
|------------|---------------------------------------|-----------------------------|------------------------------------|-----------------------|------------------------------|
| New York   | 447.3                                 | < 2.2e-16                   | 408                                | < 2.2e-16             | 7549                         |

Table 1b- Statewide COVID-19 binomial testing results for male infections that reported a significant p value and an expected probability greater than 0.50

| State      | Chi Square Statistic for Male Infection | P Value for Male Infection | n, the total number of infections |
|------------|---------------------------------------|-----------------------------|---------------------------------|
| New Jersey | 133.86                                | < 2.2e-16                   | 74361                           |
| Utah       | 17.956                                | 2.26e-05                    | 2671                            |
| South Dakota | 9.3982                          | 0.002172                    | 1311                            |

Table 1c- Statewide COVID-19 binomial testing results for male deaths that reported a significant p value and an expected probability greater than 0.50

| State          | Chi Square Statistic for Male Death | P Value for Male Death | n, the total number of deaths |
|----------------|-----------------------------------|-----------------------|------------------------------|
| Arizona        | 6.3404                            | 0.0118                | 131                          |
| Colorado       | 32.22                             | 1.377e-08             | 669                          |
| Florida        | 48.976                            | 2.591e-12             | 820                          |
| Illinois       | 23.203                            | 1.457e-06             | 944                          |
| Indiana        | 18.948                            | 1.344e-05             | 466                          |
| Michigan       | 183.22                            | 1.505e-10             | 2093                         |
| North Carolina | 18.916                            | 1.366e-05             | 131                          |
| Ohio           | 14.46                             | 0.0001432             | 389                          |
| Pennsylvania   | 12.057                            | 0.000516              | 702                          |
| Virginia       | 5.1911                            | 0.0227                | 293                          |
| Washington     | 5.5657                            | 0.01832               | 652                          |
| Wisconsin      | 4.508                             | 0.03374               | 230                          |

Using global health 5050, it was discovered that 51 countries reported either male deaths or male infections. Forty-three countries reported male infection rates. Out of those 43 countries, 12 countries were found to have a significant p value and an expected probability greater than 0.50 for both male infection and male death by COVID-19. These countries included: Thailand, The Philippines, Greece, Dominican
Republic, China, Ecuador, Peru, Mexico, Iran, Pakistan, India, and Indonesia (Table 2a). Six countries had a significant p-value and an expected probability greater than 0.50 for male infections by COVID-19. However, these countries did not report death data aggregated by sex. These countries include Japan, Singapore, Panama, Algeria, Bangladesh, and Chile (Table 2b). Twenty countries had a significant p-value and an expected probability greater than 0.50 for male deaths by COVID-19 (Table 2c). These data denote out of 43 countries reporting death frequencies for males, 32 have a frequency higher than 0.50. In Greece, Dominican Republic, The Netherlands, Thailand, and Romania the male frequency of death when compared to male infections was more than twice the frequency found for female deaths when compared to female infections. In all countries tested that reported male infections and male death frequencies, all but India and Pakistan had male death frequencies compared to male infections at least 1.5 times higher than female death frequencies when compared to female infections (Supplemental Table 1).

**Table 2a- Global COVID-19 binomial testing results for both male infections and male deaths that reported a significant p value and an expected probability greater than 0.50**

| Country       | Chi square statistic for male infection | P value for male infection | Chi square statistic for male death | P value for male death | n, the total number of deaths |
|---------------|----------------------------------------|---------------------------|------------------------------------|------------------------|------------------------------|
| Thailand      | 10.541                                 | 0.001167                  | 18.451                             | 1.743e-05              | 48                           |
| Philippines   | 23.749                                 | 1.097e-06                 | 56.635                             | 5.246e-14              | 437                          |
| Greece        | 31.306                                 | 2.204e-08                 | 26.726                             | 2.344e-07              | 116                          |
| Dominican Republic | 31.77                                | 1.736e-08                  | 79.054                             | < 2.2e-16               | 235                          |
| China         | 22.37                                  | 2.249e-06                 | 1352.96                            | < 2.2e-16              | 2114                         |
| Ecuador       | 328.15                                 | < 2.2e-16                 | 89.435                             | < 2.2e-16              | 507                          |
| Peru          | 940.32                                 | < 2.2e-16                 | 94.162                             | < 2.2e-16              | 445                          |
| Mexico        | 243.23                                 | < 2.2e-16                 | 123.75                             | < 2.2e-16              | 857                          |
| Iran          | 293.82                                 | < 2.2e-16                 | 27.637                             | 1.463e-07              | 853                          |
| Pakistan      | 2842.8                                 | < 2.2e-16                 | 52.25                              | 4.887e-13              | 209                          |
| India         | 1099.2                                 | < 2.2e-16                 | 23.488                             | 1.257e-06              | 111                          |
| Indonesia     | 308.16                                 | < 2.2e-16                 | 100.18                             | < 2.2e-16              | 773                          |

**Table 2b- Global COVID-19 binomial testing results for male infections that reported a significant p value and an expected probability greater than 0.50**

| Country | Chi square statistic for male infection | P value for male infection | n, the total number of infections |
|---------|----------------------------------------|---------------------------|----------------------------------|
| Japan   | 338                                    | < 2.2e-16                 | 10432                            |
Globally it was determined there is a significant difference in SNP occurrence for males and females for the ACE2 receptor (Welch's p-value <.001, df = 129, F=70.58). To control for potential type I errors, Benjamini & Hochberg False Discovery Rate (FDR) analysis indicated 987 pairs with significant differences (q-value <0.05). The 4 SNPs analyzed were averaged for males and females across populations. Males at higher risk for severity of disease would be males with SNPs for alternate alleles causing increased expression of the ACE2 receptor displayed in green, noted as Male1 (Figure 1, Supplemental Table 3). Males are found in all populations to have greater than 50% of the population containing the alternate alleles. Males with the
largest SNP occurrence are found in African populations as well as in the African American population, ASW. This finding may contribute to increased death seen in the United States among the African American population. In most Asian countries, only the alternate alleles are found for both male and female populations. Males with the ancestral allele, yellow, noted as Male0, (Figure 1, Supplemental Table 3), which would lead to decreased ACE2 expression, are not found in abundance globally (<50% in all populations) indicating male populations have increased expression of the ACE2 receptor. Females across some populations have only the alternate allele, in dark blue noted as Female11, which increases ACE2 expression (Figure 1, Supplemental Table 3). These populations include African populations and their descendants such as ASW, Asian populations, and African Caribbeans in Barbados, Peruvians, Gujarati Indian, and those in the United States with Mexican Ancestry. These populations may be at increased risk of death for female populations as they have females homozygous for the alternate allele. European populations and the United States population CEU, which represents European ancestry, contain females with a much higher level of heterozygosity, shown in blue, noted as Female 01, meaning they contain one ancestral and one alternate allele (Figure 1, Supplemental Table 3). Other populations with a higher proportion of heterozygous females would be African populations and Puerto Ricans, Columbians, and Peruvians. Though they contain the alternate allele, these females may be protected in a subset of their cells due to X-inactivation. Females who are homozygous for the ancestral allele (light blue, noted as Female00), meaning they do not contain an alternate allele, are found in the European populations; CEU from the United States (representing European descent), those from the United States with Mexican Ancestry and African Ancestry, Columbians, Puerto Ricans and in very small amounts the African populations (Figure 1, Supplemental Table 3). These patterns demonstrate most males are at a higher risk than their female counterparts.

**Role of single nucleotide polymorphisms in population related disease outcomes**

Worldwide, the average frequency for the alternate alleles (SNPs) for the ACE2 receptor (in pink) are seen at increased levels in most populations with highest levels in African populations including ASW of African descent and all US populations (Figure 2, Supplemental Table 4). Average means across populations show significant differences for ACE2 SNPs (Welch’s ANOVA p-value < 2.2e-16, df=25, F=596.71, Figure 3).

Benjamini & Hochberg FDR analysis found 70 significant pairs across populations (q-value <0.05). A SNP which increases expression of TMPRSS2, which is believed to cleave the viral S protein into S1 and S2 prior to virus entry into the cells is found at similar levels across the global population (shown in peach, Figure 2, Supplemental Table 4). For all populations the alternate allele was found at a frequency of at least 50% with the country having the highest frequency being Italy with 81.3%. Other populations with greater than 80% frequency for the alternate allele include Peruvians (80.2%), Esan in Nigeria (81.3%) and Southern Han Chinese (81%) (Figure 2, Supplemental Table 4). Chi square analysis across global populations did not find a significant difference in this SNP between males and females for TMPRSS2.
SNPs found across the transcription start site (TSS) in GTEx for ANPEP eQTL cause a decrease in ANPEP expression. ANPEP is believed to be a co-receptor that may assist in SARS-CoV-2 entry into cells. Statistical analysis did not indicate significant differences in prevalence across the global populations for these SNPs. This likely indicates if ANPEP is acting as a co-receptor, no population has increased protection over the others as no population has an increased prevalence for SNPs that would lower ANPEP expression. Chi-square analysis did not find significant differences for these SNPs between males and females across populations. As ANPEP is not significant across populations it is not included in Figure 2.

ENPEP; also believed to be a co-receptor for viral entry into cells does have significant differences across global populations, Welch's ANOVA (p-value = <.001, df = 25, F=2.93, Figure 4); however, Benjamini & Hochberg FDR analysis indicates no pairs are significant at a q-value < 0.05. There are thirty paired comparisons with q-values of <0.1. We still believe that ENPEP countries with increased frequencies for the alternate alleles may have important consequences as seen in purple and believed to increase ENPEP expression (Figure 2). Countries with these SNPs include Italy and Spain with the highest frequencies followed by the Utah population (European ancestry) in the United States, Great Britain, Finland and Puerto Rico (Figure 2, Supplemental Table 4). Populations from the United States with African and Mexican ancestry also contain ENPEP SNPs as well as African populations (Figure 2, Supplemental Table 4). Additionally, what may be of importance is Asian populations do not contain SNPs for ENPEP (Figure 2, Supplemental Table 4). Chi square analysis did not show a significant difference by sex across populations.

**Discussion**

ACE2 is found in a region of the X chromosome that is variable for genes escaping X inactivation. ACE2 once believed to be an escape gene displays heterogeneous sex bias across tissues with a propensity for increased expression in males in several tissues including lung, esophagus and stomach, some of the very tissues involved in COVID-19 infection [22]. Increased expression of ACE2 may exacerbate this sex bias in tissues and may increase the risk males are subjected to from COVID-19. In fact, a study investigating ACE2 expression patterns among 8 lung transplant donors found no association between ACE2 expression and smoking but did find higher ACE2 expression in male donors [10]. SNPs near the TSS of the ACE2 receptor are associated with increased expression of this protein [32-33]. Males with these SNPs may have increased expression of the ACE2 receptor in all their cells expressing ACE2 as they only have one X chromosome (Figure 5). Females with similar expression patterns to these males would be those homozygous for the SNPs, meaning they would have two X chromosomes both containing alleles for the SNPs. However, females undergoing X inactivation would reduce their amount of ACE2 receptor to the same level as that of a male who has an X chromosome containing the SNPs [22] (Figure 5). Heterozygous females, meaning they have SNPs on one X chromosome and no SNPs on the other, would not have as high level of ACE2 expression as males with the SNPs [22-23]. This is due to random X inactivation in a woman's cells. Only a percentage of her cells would have increased ACE2 expression offering a level of protection (Figure 5). Women who are homozygous for no SNPs would have less ACE2 expression overall as they would still undergo X inactivation leading to cells expressing one X chromosome. This would be
comparable to males who do not have SNPs. In all populations looked at in this study the frequencies for males with SNPs were greater than 50%. Females who were homozygous for the SNPs were found at much lower frequencies. Lung cells of males have been shown to have a higher number of ACE2 receptors than females which leads to increased infection [7,10]. This increased number of ACE2 receptors may be directly related to increased expression of the ACE2 gene as more than 50% of males across the globe contain SNPs associated with increased expression of the ACE2 gene. It has been shown that polymorphisms in the ACE2 gene correlate with increased risk for hypertension; therefore, it is not unreasonable to believe that ACE2 polymorphisms could also affect one's risk for COVID-19 infection [29]. As ACE2 expression may be increased and is seen in lungs, heart, kidney, ileum, brain, esophagus, colon, nasal goblet and ciliated cells these tissues could be at risk for attack by the virus. Additionally, ACE2 plays an important role in the homeostasis of the renin-angiotensin system (RAS) which is important for regulating heart, kidney and lungs both physiologically and pathologically. When SARS-CoV-2 binds ACE2 there is systematic deprivation of ACE2 to the tissues leading to an increase in vasoconstriction, inflammation and fibrosis [7].

To investigate what makes some populations or individuals of an ancestry more susceptible to COVID-19, one needs to consider the genetic players contributing to infection and severity of disease, of which ACE2 is only one. TMPRSS2 was found to be highly co-expressed with ACE2 in nasal goblet and ciliated cells which could serve as locations for initial infection or as reservoirs for infectivity between individuals [9]. This co-expression is found in esophagus, ileum, and colon as well, which may lead to other routes of infection [9,21]. Populations with larger frequencies of SNPs increasing ACE2 and TMPRSS2 expression could increase susceptibility to infection as well as severity of disease across multiple tissues. ENPEP, like ACE2 plays an important role within the RAS system and with hypertensive conditions as ENPEP has higher expression levels in the kidneys. Cardiac and nephrotic complications have been reported with SARS-CoV-2 patients. In fact, COVID-19 associated nephritis may serve as a marker of disease severity and capillary leak syndrome; a predictor of fluid overload, respiratory failure, and death [34]. Additionally, ENPEP receptors are suggested to play an important role in the RAS system in the brain [12,16,17]. Patients are exhibiting symptoms of brain clots and stroke related to COVID-19 infection even in younger patients. The Mount Sinai Health System treated 5 COVID-19 patients all under 50 years of age for large vessel stroke over a two-week period. This same health system only reported 0.73 patients over a two-week period for the last 12 months [35].

In total, populations that have increased frequencies of SNPs may have increased expression of all three of these formidable proteins for COVID-19 initiation and propagation contributing to this insidious disease. Populations with the highest frequencies of SNPs for ACE2, TMPRSS2 and ENPEP include all populations of the US, Italy, Finland, Spain, United Kingdom, Puerto Rico, Nigeria, Kenya, and Barbados. Countries with the highest death frequency compared to infection calculated in this study include Italy, Sweden, The Netherlands, Belgium, Hungary, Mexico, Spain, Indonesia, and Northern Ireland. The United Kingdom and United States did not provide values for this calculation. Asian countries do not have any SNPs investigated in this study for ENPEP and have some of the lowest death frequencies calculated. Lack of SNPs for ENPEP may protect individuals in those populations from the higher death rates seen with other
populations. This study may serve as a building block for future genetic studies to further investigate the role these variants as well as other variants may play in the severity of this disease.

Countries that may be winning the battle against COVID-19 may be doing so because of the genetic make-up of their people. Recent challenges found while investigating treatments for COVID-19 may be due to ancestral and sex differences leading to differential expression of these important proteins, thus convoluting treatment effect. The susceptibility of individuals to this disease may be due to each person's coordinated concert of genetic milieu making treatment efficacy challenging without taking the potential for personalized medicine into consideration.

**Methods**

**Statistical Testing for male infection rates and death rates**

A binomial test was used to compare the probability of males versus females who either contracted COVID-19 or died from the disease. R (version 3.6.3) was used to perform the binomial test [36]. In this code, $x$ equaled the number of infected males or the males who died from COVID-19, where $n$ equaled the total number of infections or deaths from COVID-19 which included both females and males. The expected $P$ equaled 0.50 due to the United States and global sex ratio being close to 50%. Therefore, it was expected that males would account for 50% of COVID-19 deaths or infections. Yates correction for continuity was not applied to this statistical test, which was indicated by correct equal FALSE. While Yates correction for continuity adjusts the chi square test for populations smaller than 5, this test would only apply to Wyoming. However, Yates correction of continuity did not change the p value for Wyoming. The p-value for this study was 0.05. Any state or country that had a p value lower than .05 with an expected probability greater than 0.5 was considered statistically significant.

The data collection for states was provided by the department of public health for each state and was documented in the Supplemental Methods Table 5. In this study, 47 states reported COVID-19 infection values aggregated by sex, but only 24 states reported COVID-19 deaths by sex. Three states—Hawaii, Nebraska, and Connecticut—did not report infection or death rates of COVID-19 and were disregarded from this study. It should be noted that some states or countries provided unknown categories. These values were excluded from the study as the focus was to compare males versus females. These changes are listed in Supplemental table 5. For some states, the department of public health reported the deaths and infection rates in percentages. To find the number of males who died from COVID-19, the percentage was multiplied by the total number of deaths from COVID-19 in that state. Additionally, to find the number of males infected with COVID-19, the percentage was multiplied by the total number of infections from COVID-19 in that state. Once the number of male deaths or infections was recorded, the binomial tested was used to determine the p value and chi square statistic (Supplemental Table 5).

The same binomial test was used for countries to compare male deaths or male infections versus females. The data collection for countries was provided by Global Health 50/50 and the sources for each country were documented in the Supplemental Tables (Supplemental Table 6a, Supplemental Table 6b). In this
study, a total of 51 countries reported either deaths or infections separated by sex. Just like some states, some countries documented the data in percentages. For male infections, the total number of infections was multiplied by the percentage of males infected in that country. For male deaths, the total number of deaths was multiplied by the percentage of males that died in that country. If the countries reported unknown values, these values were subtracted from the total and are documented in the Supplemental Tables. After gathering the data, the binomial test was conducted in the programming software R (version 3.6.3) to analyze COVID-19 deaths and infections by sex in each country [36].

Lastly the frequency of COVID-19 deaths for each country was collected. The frequency was collected by taking the total deaths of COVID-19 divided by the total infections of COVID-19 for that country. The frequency was also done for males and females by taking males deaths of COVID-19 divided by male infections of COVID-19 or female deaths of COVID-19 divided by female infection of COVID-19. Out of the 51 countries used in the binomial test, 48 were used for the total frequency test because England and Wales, United States, and Scotland did not report total infections. However, there were an additional 19 countries added for total frequency. This is documented in the Supplemental Tables. For male and female frequency, only 35 countries reported both male infection and male deaths and these countries are noted. Lastly, it should be noted that the Supplemental Tables include date accessed, sources for both state and country, and any additional information used in the study. Global 50/50 [37] was the website used that contained all data across countries for COVID-19. However, the site provided where the original information was obtained, which was documented as the primary source for that country. All 51 countries listed in Supplementary Table 6a were used for binomial testing. However, for total frequencies three countries were excluded: England and Wales, United States of America, and Scotland. For frequency data, if the country contained unknowns, then frequency data also contained unknowns.

Selection and Use of Single Nucleotide Polymorphisms (SNPs)

SNPs chosen to represent the genes in this study were close to the TSS as the TSS is important in gene regulation [38,39,40]. Additionally, SNPs chosen were found in eQTL associations across multiple tissues. SNPs (rs4646120, rs1978124, rs5934250, rs6632680) were chosen for the ACE2 receptor from GTEx eQTL data [32,33]. Upregulation of the ACE2 receptor was found with the G allele, the C allele, the G allele, and the C allele respectively (Supplemental Figure 1a). SNPs (rs113104244, rs78606958, rs1552456) were chosen for the ENPEP receptor from GTEx eQTL data with eQTL associations in the lung [32,33]. Upregulation of the ENPEP receptor was found with the A allele, the G allele, and the G allele respectively (Supplemental Figure 1b). SNPs (rs12442778, rs11635469, rs73478036, rs28565347) were chosen for the ANPEP receptor from GTEx eQTL data [32,33]. Downregulation of the ANPEP receptor was found with the G allele, the G allele, the G allele, and the T allele respectively (Supplemental Figure 1c). For TMPRSS2, eQTL was only found in testes for GTEx associations and did not meet our criteria for eQTL associations across multiple tissues. Therefore, the eQTL used to choose the SNP for analysis is from Sudamant et al., 2015 [41]. To confirm SNP choices, principle component analysis was utilized to determine possible linkage
disequilibrium for SNPs chosen [36]. PCA demonstrated SNPs chosen would be expected to inherit together within populations (Supplementary Figure 2). As only one SNP was chosen for TMPRSS2 ANOVA testing was not carried out; however, frequency for populations and Chi Square for males and females within populations was calculated [36].

SNPS chosen from GTEx data were then searched in Thousand Genomes Database, NCBI [42-44]. Frequencies for alleles which affect gene expression were calculated (in most cases the alternate allele frequency). These frequencies were used for subsequent statistical calculations. Welch’s ANOVA testing was used for ACE2 SNPs for males and females as well as ACE2 and ENPEP SNPs across populations as data were not normally distributed and homogeneity of variance was violated. R (version 3.6.3) was used for all statistical testing including Benjamini & Hochberg FDR [36, 45]

ARC GIS

Spatial analysis was conducted in ArcMap 10.4.1. All maps were made using the GCS_WGS_1984 map projection in decimal degrees. All base maps used were the light gray canvas base maps from ESRI. Symbology of the average frequencies of the ACE2 receptor by sex and average Eqtl of ACE2, ANPEP, ENPEP, and TMPRSS2 SNPSs was done using the pie chart options in ArcMap. Colorblind friendly color schemes were chosen using the website colorbrewer.com. From ColorBrewer site, The original ColorBrewer (v1.0) was funded by the NSF Digital Government program during 2001-02, and was designed at the GeoVISTA Center at Penn State (National Science Foundation Grant No. 9983451, 9983459, 9983461). The design and rebuilding of this new version (v2.0) was donated by Axis Maps LLC, winter 2009 and updated in 2013.

Declarations

Acknowledgements

Jason C. Dobrinski for his assistance with Figure 5.

Additional Information

Data Availability Statement

The datasets analyzed for SNP frequency are publicly available in the NCBI Thousand Genomes repository https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/ and GTEx data repository, https://www.gtexportal.org/home/locusBrowserPage/ for eQTL data.

The datasets utilized for COVID-19 death and infection data from the United States are publicly available from each state’s Department of Public Health. This location is also listed by state in Supplementary Results Table 5.
The datasets utilized for global COVID-19 death and infection data was obtained and is publicly available from Global Health 50/50 https://globalhealth5050.org/covid19/.

Global Health 50/50 reports where each country’s data was obtained. We have included this information for each country from Global Health 50/50 in Supplementary Results Table 6.

Competing Interests Statement

The authors declare no competing interests.

Authorship contributions

All authors contributed extensively to the work presented in this paper.

K.C., M.H., and F.B.M. contributed equally to this work. K.C. found all published state and country data to carry out the binomial testing and death frequencies. K.C. created supplemental tables 1,2,3,7 and 8. M.H. carried out the average SNP frequency calculations and created the global maps. M.H. also created supplemental tables 4,5 and 6. F.B.M. carried out the literature search and wrote the introduction. B.F.F. oversaw the creation of all the maps and wrote the figure legends. B.F.F. also assisted in editing of the paper. K.P.D designed the study, chose the SNPs based on eQTL studies (GTEx). K.P.D. carried out all statistical analysis other than binomial testing and death frequencies. K.P.D created figures 3,4,5. K.P.D wrote the main paper and Supplementary information including online methods. All authors discussed the results and discussion and commented on the manuscript at all stages.

First authors Keyana Cardoza, Magen Hoch, and Frederic B. Montz contributed equally to this work.

References

1. John Hopkins University of Medicine. (2020, April 19). Coronavirus Resource John Hopkins University. Retrieved from: https://coronavirus.jhu.edu/map.html

2. World Health Organization. (2020). COVID-19 weekly surveillance report. Retrieved from: https://infogram.com/1prl06x5qnekmwtg10l6rd3qe3bmdpjdj3g76

3. Fehr, A. R. & Perlman, S. Coronaviruses: Methods and Protocols 1–23 (Springer New York, 2015). doi:10.1007/978-1-4939-2438-7_1

4. Lan, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature (2020). doi:10.1038/s41586-020-2180-5

5. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. Med. 26, 450–452 (2020).

6. Zhang, Y. Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell 181, 223–227 (2020).
7. Li, Y., Zhou, W., Yang, L. & You, R. Physiological and pathological regulation of ACE2, the SARS-CoV-2 receptor. *Res.* **157**, (2020).

8. Qi, F., Qian, S., Zhang, S. & Zhang, Z. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. *Biophys. Res. Commun.* (2020). doi:10.1016/j.bbrc.2020.03.044

9. Sungnak, *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* (2020). doi:10.1038/s41591-020-0868-6

10. Zhao, Y. *et al.* Single-cell RNA expression profiling of ACE2, the receptor of SARS-CoV-2. Preprint at [https://doi.org/10.1101/2020.01.26.919985](https://doi.org/10.1101/2020.01.26.919985) (2020).

11. Chen, Y., Shan, K. & Qian, W. Asians do not exhibit elevated expression or unique genetic polymorphisms for ACE2, the cell-entry receptor of SARS-CoV-2. Preprint at doi:10.20944/preprints202002.0258.v2 (2020).

12. Holmes, R. S., Spradling Reeves, K. D. & Cox, L. A. Mammalian Glutamyl Aminopeptidase Genes (ENPEP) and Proteins: Comparative Studies of a Major Contributor to Arterial Hypertension. *Data Mining Genomics Proteomics* **08**, (2017).

13. Surendran, P. *et al.* Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Genet.* **48**, 1151–1161 (2016). https://doi.org/10.1038/ng.3654

14. Mizutani, S. *et al.* New insights into the importance of aminopeptidase A in hypertension. *Heart Fail Rev.* **13**, 273–284 (2008).

15. Moore, N. *et al.* Renin gene polymorphisms and haplotypes, blood pressure, and responses to renin-angiotensin system inhibition. **50**, 340–347 (2007).

16. De Mota, N. *et al.* Human brain aminopeptidase A: Biochemical properties and distribution in brain nuclei. *Neurochem.* **106**, 416–428 (2008).

17. Speth, R. C. & Karamyan, V. T. The significance of brain aminopeptidases in the regulation of the actions of angiotensin peptides in the *Heart Fail. Rev.* **13**, 299–309 (2008).

18. Baig, A. M., Khaleeq, A., Ali, U. & Syeda, H. Evidence of the COVID-19 Virus Targeting the CNS: Tissue Distribution, Host-Virus Interaction, and Proposed Neurotropic Mechanisms. *ACS Chem Neurosci.* (2020). doi:10.1021/acscchemneuro.0c00122

19. Yeager, C. L. *et al.* Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature* **357**, 420–422 (1992).

20. Huang, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
21. Lukassen, S. et al. SARS-CoV-2 receptor ACE 2 and TMPRSS 2 are primarily expressed in bronchial transient secretory cells. *EMBO* (2020). doi:10.15252/embj.20105114

22. Tukiainen, T. et al. Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248 (2017).

23. Franco, B. & Ballabio, A. X-inactivation and human disease: X-linked dominant male-lethal disorders. *Opin. Genet.* **16**, 254–259 (2006).

24. Deng, X., Berletch, J. B., Nguyen, D. K. & Distèche, C. M. X chromosome regulation: Diverse patterns in development, tissues and *Nat. Rev. Genet.* **15**, 367–378 (2014).

25. Shaikh, H. et al. High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Res.* **19**, 1682–1690 (2009).

26. Krausz, C. et al. High Resolution X Chromosome-Specific Array-CGH Detects New CNVs in Infertile Males. *PLoS ONE* **7**, (2012).

27. Lee, H. et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).

28. Jiao, Y. et al. Single nucleotide polymorphisms predict symptom severity of autism spectrum disorder. *Autism Dev. Disord.* **42**, 971–983 (2012).

29. Lu, N. et al. ACE2 gene polymorphism and essential hypertension: An updated meta-analysis involving 11,051 subjects. *Biol. Rep.* **39**, 6581–6589 (2012).

30. Stawiski, E. W. et al. Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility. Preprint from https://doi.org/10.1101/2020.04.07.024752

31. Cao, Y. et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**, (2020).

32. The GTEx The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015)

33. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

34. Gross, O., Moerer, O., Weber, M., Huber, T. B. & Schieithauer, S. COVID-19-associated nephritis: early warning for disease severity and complications? (2020). https://doi.org/10.1016/S0140-6736(20)31041-2

35. Oxley, T. J. et al. Large-Vessel Stroke as a Presenting Feature of Covid-19 in the Young. *N. Engl. J. Med.* (2020). doi:10.1056/NEJMc2009787
36. R Core Team (2020). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

37. The Global Health 50/50 Report 2020: power, privilege and priorities. https://globalhealth5050.org/2020report/. Date: March, Date accessed: March 25, 2020.

38. Ngoc, L. V., Cassidy, C. J., Huang, C. Y., Duttke, S. H. C. & Kadonaga, J. T. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. Genes Dev. 31, 6–11 (2017).

39. Kugel, F. & Goodrich, J. A. Finding the start site: Redefining the human initiator element. Genes Dev. 31, 1–2 (2017).

40. Umarov, R. et al. Promoter analysis and prediction in the human genome using sequence-based deep learning models. 35, 2730–2737 (2019).

41. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. Nature 526, 75–81 (2015).

42. The 1000 Genomes Project Consortium, Auton, A. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).

43. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information. NAR 35, (2007).

44. Clarke, L., et al. The International Genome Sample Resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data, NAR 45, (2017) https://doi.org/10.1093/nar/gkw829

45. Benjamini, , and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. STA.T SOC. B, 57, (1995).

Figures
Average frequencies of the ACE2 receptor by sex data. Male0 populations are males with the ancestral allele that decreased ACE2 expression and Male1 populations are males with SNPs for alternate alleles causing increased ACE2 receptors. Female00 populations are females that are homozygous for the ancestral allele (decreased ACE2 expression), Female01 populations are females that are heterozygous, containing one ancestral and one alternate allele, and Female11 are females who are homozygous for the alternate alleles (increased ACE2 expression). Globally, the frequency of occurrence of males with SNPs for the alternate alleles was higher than males or females with the ancestral allele in all populations. The mean frequency of the ACE2 receptor was substantially higher for males (0.787 ± 0.142) than females (0.599 ± 0.247). Frequency of alternate alleles causing increased ACE2 receptors in males ranged from 0.51 – 1. Whereas the frequency of females for homozygous alternate alleles ranged from 0.14 – 0.98). Furthermore, the CDX population had the highest male frequency (1) and the JPT population had the highest female frequency (0.98) for the alternate allele. The frequency for the ancestral allele in males was lowest in the KHV population (0.005) and was highest in the TSI population (0.486). The frequency for homozygous ancestral allele in females was lowest in the KHV, CHS, JPT, and PEL (0) populations and
highest in the CEU population (0.32). The frequency for heterozygous allele in females was lowest in the JPT (0.015) population and was highest in the FIN (0.610) population. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.
Averages of TMPRSS2, ACE2, and ENPEP SNPs. Globally there was an increase of frequency for ACE2 SNPs with the CEU population having the lowest frequency (0.48) and JPT having the highest frequency (0.988). Furthermore, the BEB, ITU, GWD, LWK, MSL, ACB, PEL, CDX, CHB, KHV, JPT, and CHS populations all have frequencies greater than 0.8 for ACE2 SNPs. The lowest frequency (0.5) for TMPRSS2 SNPs was in the YRI and BEB populations and the highest frequency was (0.813) was in the TSI and CEU populations. Frequency of ENPEP SNPs was substantially lower in all populations. Populations CDX, JPT, KHV, CHS, and CHB all had no frequency for ENPEP SNPs. The highest frequency (0.140) was in the IBS population. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.
Figure 3

Barplot of combined mean frequencies for SNPs for ACE2 receptor across all populations. Error bars represent standard error calculated in R (Version 3.6.3). The smallest frequencies <50% were found in populations CEU (US populations of European descent) and TSI (Italy) with the largest frequencies >90% occurring for CDX, CHB, KHV, CHS, JPT (Four Chinese populations and the Japanese population).
Figure 4

Barplot of combined mean frequencies for SNPs for ENPEP receptor across all populations. Error bars represent standard error calculated in R (Version 3.6.3). The smallest frequencies 0% were found in populations CDX, CHB, KHV, CHS, JPT (Four Chinese populations and the Japanese population). These are the same populations found to have the largest ACE2 SNP frequency. The largest frequencies >1% are found for populations PUR, FIN, GBR, CEU, TSI, IBS (Puerto Rico, Finland, Great Britain, US population with European ancestry, Italy, and Spain respectively).
a. Male
Ancestral alleles containing no SNPs

X-inactivated

b. Alternate alleles containing SNPs

Every cell overexpresses ACE2

X-inactivated cells do not overexpress ACE2

X-inactivated cells overexpress ACE2

c. Viruses break out of host cell; infect easily as all male cells overexpress ACE2

Viruses break out of host cell; does not infect female cells as easily; not every cell overexpresses ACE2
Figure 5

Representation of X chromosomes for males and females with and without SNPs. Row a. Purple bands represent ACE2 gene with no alternate alleles (no SNPs) to increase gene expression. Females would undergo X inactivation of one X chromosome in all female cells making gene dosage in male and female cells similar. Row b. ACE2 gene with alternate alleles (with SNPs) are indicated in orange. Male cells contain an X chromosome with SNPs increasing ACE2 gene expression in all male cells. Some females contain one X chromosome with SNPs (orange) and one without SNPs (purple). X inactivation would be random in female cells meaning some cells would have X chromosome inactivation for the chromosome that does not contain SNPs (purple). Those cells would have increased expression of the ACE2 receptor similar to males in row b. Other cells would have X chromosome inactivation for the X chromosome that does contain SNPs (orange) and would therefore not have increased expression of the ACE2 receptor. As some of the female's cells will not have increased ACE2 expression they may have a level of protection. Row c. As viruses break out of their host cells and move to other cells for further infection all male cells will be overexpressing ACE2 as they contain SNPs and may be easily infected. Some female cells will overexpress ACE2 while others will not. This may limit the ability of the virus to infect new female host cells.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryresultsfinalSR.pdf