Analysis of COVID-19 genetic risk susceptibility using UK Biobank SNP genotype data

Taewan Goo, Kyulhee Han and Catherine Apio

Interdisciplinary Program in Bioinformatics,
Seoul National University,
Gwanak-gu, Seoul, South Korea
Email: gootec92@snu.ac.kr
Email: hgh1031@snu.ac.kr
Email: 2019-20240@snu.ac.kr

Taesung Park*

Department of Statistics,
Seoul National University,
Gwanak-gu, Seoul, South Korea
Email: tspark@stats.snu.ac.kr
*Corresponding author

Abstract: The coronavirus disease 2019 (COVID-19) has become a global pandemic. Here, we performed a study on host susceptibility to COVID-19 infection using COVID-19 test results and genomic data released by UK Biobank until early October of 2020. The data consisted of 27,713 samples including 2740 positive cases. We employed genome-wide association study, gene-level association and pathway analyses using common and rare variants. Among these analyses, only pathway analysis based on rare variants found seven significant pathways. Among them, the JAK-STAT pathway and glycolipid biosynthesis pathway have been reported to be associated with a viral infection, especially COVID-19 infection. Further, we found new pathways that were not previously reported, including pathways related to cellular signalling like NLR signalling pathway. Additional experiments and studies of these pathways may unveil the pathophysiological of COVID-19 and identify highly susceptible groups.

Keywords: COVID-19; GWAS; host genetics; infection susceptibility; pathway analysis.

Reference to this paper should be made as follows: Goo, T., Han, K., Apio, C. and Park, T. (2021) ‘Analysis of COVID-19 genetic risk susceptibility using UK Biobank SNP genotype data’, Int. J. Data Mining and Bioinformatics, Vol. 25, Nos. 1/2, pp.1–16.

Biographical notes: Taewan Goo is a PhD student in Interdisciplinary Program in Bioinformatics, Seoul National University.

Kyulhee Han is a PhD student in Interdisciplinary Program in Bioinformatics, Seoul National University.

Catherine Apio is MSc student in Interdisciplinary Program in Bioinformatics, Seoul National University.
1 Introduction

In the early stage of the spread of coronavirus disease 2019 (COVID-19), the World Health Organisation (WHO) declared COVID-19 a global pandemic on 11 March 2020 (https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020). As of 24 February 2021, there were a total of 112,752,364 COVID-19 confirmed cases and 2,498,659 deaths worldwide (https://www.worldometers.info/coronavirus/). Efforts in developing vaccines and other treatment methods by different countries and institutions, to overcome the pandemic situation are being made. Therefore, many organisations and institutions have been collecting host genomes and clinical information and started research on COVID-19 host genetics (https://www.covid19hg.org/about#overview). An example of such an institution is the UK Biobank (UKB).

The UKB is a large British organisation for collecting and providing materials for this cohort study. UKB has been collecting various medical records and genomic data of about 500,000 people since 2006. Recently, UKB started working with medical institutions to quickly collecting and refining COVID-19 tests for the UKB cohort. For the mitigation and suppression of the COVID-19 pandemic, the UK government implemented different inspection policies over time, depending on the trend of prevalence and ability to accept inspections. For example, the UK government tested only hospitalised patients with respiratory symptoms until early April, but gradually expanded the target of testing to those with a contact history with a COVID-19 patient. In this study, we performed analyses using UKB data released in early October 2020.

Although infectious diseases are caused by external pathogens, the biological characteristics of the infected host are always closely related to the signs and symptoms or susceptibility of the disease. Through host genetics, associations between host genomes and susceptibility to infection by a pathogen could be revealed. For example, Nguyen et al. (2020) reported that different types of HLA show varying affinities to SARS-CoV-2 proteome suggesting that the affinities of HLA could be associated with vulnerability to the infection. It was also reported that TMPRSS2 and ACE2 DNA polymorphisms were likely associated with genetic susceptibility to COVID-19 (Hou et al., 2020). Individual genetic variation may help explain different immune responses to a virus across a population and help identify highly susceptible groups, to the disease. This allows for a more detailed plan for setting up infection-related policies such as prioritising vaccination for high-risk HLA individuals to COVID-19 (Nguyen et al., 2020).
Analysis of COVID-19 genetic risk susceptibility

However, most of the current host genetics studies have only focused on revealing associations between a single variant such as a Single Nucleotide Polymorphism (SNP) and infection (single-allele study designs), which makes it difficult to be applied to the study of rare variants with Minor Allele Frequency (MAF) smaller than 1%. Factors such as locus heterogeneity, epistasis and multiple genes conferring small effects contribute to the complexity of the genetic models underlying phenotype expression. Thus, many biologically meaningful associations having lower effect sizes at individual genes are overlooked, making it difficult to separate true associations from a sea of false-positive associations. Organising these individual SNPs into biologically meaningful groups to look at the overall effects of minor perturbations to genes and pathways is desirable (White et al., 2019). These gene-level and pathway-based approaches provide researchers with insight into the functional foundations of the phenotype being studied and allows testing of various genetic scenarios. Thus, gene- and pathway-level analyses have been suggested as an alternative method.

In this study, genetic variants were divided into two groups for analysis: 1) common and less common variants and 2) rare variants. Common variants are variants with MAF exceeding 5%. Rare variants, on the other hand, have less than 1% MAF (Bycroft et al., 2018). Less common variants were defined as variations with MAF between 1% and 5%. Owing to the low MAF of rare variants, large sample sizes are always required to achieve statistically significant results. Therefore, different methods have been developed for the analysis of rare variants, different from those employed in the analysis of common variants. Also, COVID-19 infection PCR test results from UKB was used, consisting of 27,713 samples with 2470 positive cases, tested in early October.

Firstly, we performed Genome-Wide Association Studies (GWAS), gene-level association and pathway analyses to determine the association between the host genome and COVID-19 infection susceptibility. The GWAS analysis was performed with only common and less common variants.

Secondly, we performed gene-level association tests. In this analysis, common and less common variants were analysed with PLINK (v1.90) set-based test and rare variants using Sequence Kernel Association Test (SKAT) and Optimised SKAT (SKAT-O) (Hoh et al., 2001; Lee et al., 2012; Wu et al., 2011). PLINK’s set-based test is proper for large-scale candidate gene studies with a permutation test. SKAT is the method that reveals the association between gene and phenotype based on multiple regression of rare variants, while SKAT-O is the optimised version of SKAT. SKAT-O maximises statistical power by optimally combining the burden tests and SKAT.

Lastly, we performed pathway analysis for rare variants, common and less common variants. We used GSA-SNP2 and Multi-marker Analysis of GenoMic Annotation (MAGMA) methods for pathway analysis of common and less common variants (De Leeuw et al., 2015; Nguyen and Nam, 2017). Pathway-based approaches using HierArchical components of collapsed RAre variants Of High-throughput sequencing data (PHARAOH), SKAT and SKAT-O methods were used for the pathway analysis of rare variants (Lee et al., 2016). Among the pathway analysis methods used for common and less common variants, GSA-SNP2 evaluates the association between pathways and phenotype based on GWAS results, while MAGMA requires Principal Component Analysis (PCA) to incorporate large numbers of variants into the gene level. Lastly, PHARAOH considers the hierarchical structure of the biological function of genes and pathways. PHARAOH also analyses entire pathways simultaneously and assumes that biomarkers have linear relationships with the phenotype of interest.
As a result, we identified significant genes and pathways associated with COVID-19 susceptibility. Some of the significant genes and pathways were previously reported in several types of research and others were not. Now, we hope our results can help in the understanding of the infection process of COVID-19 and suggest novel gene and pathways related to COVID-19 susceptibility. Furthermore, the results of the analysis could be used as the first step for identifying groups in urgent need of vaccines or preventing infection of diseases, such as studies of existing high infection risk groups.

2 Materials and methods

2.1 UKB data

UKB is a longitudinal prospective cohort study to improve public health through investigating the causes, treatment and prevention of diseases of middle and older age (Galinsky et al., 2016). The cohort included about 500,000 men and women aged 40–69 in 2006–2010. Since the early stage of the COVID-19 pandemics, UKB has been receiving test data from Public Health England (PHE). The PCR test was conducted on biospecimen obtained from the nose or neck. By early October, the number of samples tested for COVID-19 infection was 27,713 out of the total cohort. To control for multiple tests performed on an individual, we considered a person was positive if the sample were found as positive at least once.

Demographic variables such as age and sex were used in GWAS. Table 1 shows the clinical characteristics of the samples. Samples in the data were collected from various ethnicities. Therefore, the first five Principal Components (PCs) were used to adjust the variation of ethnicity as suggested in previous research (Kent, 2011). Genome-wide genotyping was performed using the UK Biobank Axiom Array. The array provides information on approximately 850,000 genetic variants and mainly contains variations that are likely to be associated with a disease. Genetic variants are divided into common (MAF ≥5%), less common (1–5% MAF), and rare variants (MAF<1%) depending on the minor allele frequency of the sample. Further details of genomic data can be found in the UK Biobank Axiom Array Content Summary (The UK Biobank, 2014).

<table>
<thead>
<tr>
<th>Data</th>
<th>Category</th>
<th>Negative (n = 24,973)</th>
<th>Positive (n = 2,740)</th>
<th>Sum (n = 27,713)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>11,867</td>
<td>1411</td>
<td>13,278</td>
<td>p = 8.296E-05 (χ² test)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>13,106</td>
<td>1329</td>
<td>14,435</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>40–49</td>
<td>4596</td>
<td>985</td>
<td>5581</td>
<td>p &lt; 2.2E-16 (two sample t-test)</td>
</tr>
<tr>
<td></td>
<td>50–59</td>
<td>7190</td>
<td>801</td>
<td>7981</td>
<td></td>
</tr>
<tr>
<td></td>
<td>60–69</td>
<td>12,999</td>
<td>1,038</td>
<td>14,037</td>
<td></td>
</tr>
<tr>
<td></td>
<td>70–79</td>
<td>188</td>
<td>16</td>
<td>204</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td>British</td>
<td>21,841</td>
<td>2302</td>
<td>24,143</td>
<td>p &lt; 2.2E-16 (χ² test)</td>
</tr>
<tr>
<td></td>
<td>Irish</td>
<td>694</td>
<td>75</td>
<td>769</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Any other with white</td>
<td>790</td>
<td>65</td>
<td>855</td>
<td></td>
</tr>
</tbody>
</table>
2.2 Genotyping data pre-processing

Genotyping data measurement and Quality Control (QC) were performed by UKB. In addition, additional marker-based QC and individual-based QC were applied according to Liberzon et al. (2011). This study suggested only samples with a missing rate of less than 10% of the subject were included and genetic variants with a missing rate of less than 10% of the marker were included. All pre-processing steps were performed using PLINK (v1.90b) (https://zzz.bwh.harvard.edu/plink/).

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used for pathway analysis (Kanehisa and Goto, 2000). All the genetic variants are mapped to gene symbols and gene ID. Then, gene symbols were mapped to the KEGG pathway. The KEGG pathway file can be downloaded from MsigDB (Timmins et al., 2020).

2.3 GWAS using common and less common variants

In this analysis, only variants with a \( p \)-value > 10\(^{-6} \) from Hardy Weinberg Equilibrium test and MAF > 0.01 were used. The total number of SNPs used for analysis was 577,605. If the COVID-19 test result was positive, the sample was considered as a case and if not as a control. Univariate logistic regression for the GWAS analysis was performed using PLINK (v1.90b) (Chang et al., 2015; Purcell and Chang, 2015).

2.4 Gene-level and pathway analysis

2.4.1 Analysis of common and less common variants

Gene-level analysis for common and less common variants was performed using PLINK set-based analysis and MAGMA. As suggested by Hoh et al. (2001), PLINK (https://zzz.bwh.harvard.edu/plink/) set-based test is a method for analysing gene or specific set of SNPs by defining set-level summary statistic as individual p-values of each SNPs. Before the PLINK set-based test, we performed GWAS for individual common and less common variants. Then, PLINK set-based test was performed using the \( p \)-values obtained from the above GWAS.

On the other hand, MAGMA performs gene and pathway analysis using multiple linear regression after Principal Components (PC) analysis for each gene. F-test \( p \)-value of each gene was obtained to determine significantly associated genes with COVID-19 infection susceptibility (De Leeuw et al., 2015). MAGMA command-line tool for Linux

---

Table 1  Clinical characteristics of metabolic syndrome-related traits (continued)

<table>
<thead>
<tr>
<th>Data Category</th>
<th>Data Category</th>
<th>Negative ((n = 24,973))</th>
<th>Positive ((n = 2,740))</th>
<th>Sum ((n = 27,713))</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnicity</td>
<td>Indian</td>
<td>384</td>
<td>64</td>
<td>448</td>
<td>( p &lt; 2.2 \times 10^{-16} ) (( \chi^2 ) test)</td>
</tr>
<tr>
<td></td>
<td>Pakistani</td>
<td>103</td>
<td>36</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Caribbean</td>
<td>239</td>
<td>48</td>
<td>287</td>
<td></td>
</tr>
<tr>
<td></td>
<td>African</td>
<td>213</td>
<td>51</td>
<td>264</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>673</td>
<td>97</td>
<td>770</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>36</td>
<td>2</td>
<td>38</td>
<td></td>
</tr>
</tbody>
</table>
was downloaded from the website of the Department of Complex Trait Genetics (CTG) Lab, Centre for Neurogenomics and Cognitive Research (CNCR) (https://ctg.cnrc.nl/software/magma). We employed the gene annotation function of MAGMA to match SNPs to genes based on their positions. From the annotation step, we obtained an annotated set file, which was then used in gene analysis.

In the pathway analysis for common and less common variants, GSA-SNP2 and MAGMA were used. First, GSA-SNP2 was performed using the \( p \)-values of each of the SNPs obtained from the above GWAS to evaluate the association of genetic variants with pathways. Yoon et al. (2018) introduced that GSA-SNP2 provides high power, decent type I error control, and efficient computing. Windows-based Graphical User Interface (GUI) of GSA-SNP2 (https://sites.google.com/view/gasasn2) was employed in this analysis (Nguyen and Nam, 2017). Each SNP was mapped to genes and KEGG pathways based on the position of the SNPs, by the annotation method of GSA-SNP2. After, we performed enrichment analysis for all variants.

Secondly, MAGMA pathway analysis was performed using the results of the MAGMA gene-level analysis. The \( F \)-test \( p \)-values of each gene were obtained from MAGMA gene-level analysis. These \( p \)-values of genes were used as inputs of the self-contained test to measure the significance of each pathway after \( z \)-transformation.

### 2.4.2 Analysis of rare variants

For the analysis of rare variants, only genetic variants with MAF less than 0.01 were selected. The number of rare variants used in the analysis was 58,564 rare variants. Because of low MAF, it is difficult to reveal associations between each rare variant and phenotype of interest through logistic regression. To deal with this problem, we performed gene-level and pathway analysis using SKAT and SKAT-O, two methods for gene-level association analysis of rare variants (Lee et al., 2012; Wu et al., 2011). Because these methods required defining sets of variants by regional or biological function, we grouped the rare variants based on their positions in genes. For SKAT and SKAT-O, we used the gene list that was generated by the annotation step of MAGMA which assigns each SNP to a gene if the genomic location of SNP was inside the region of a specific gene. We performed the analysis using the \( R \) package (“SKAT”).

In addition, three methods were employed in the pathway analysis of rare variants. First, SKAT and SKAT-O used a set of SNPs for pathway level. Thus, variants were mapped into KEGG pathways and performed the analysis as previously we explained. The last method used was PHARAOH that uses hierarchical structures of collapsed rare variants and biological pathway information. Lee et al. (2016) suggested that PHARAOH can reflect the biological functional structure of a gene to a pathway and takes in to account the multicollinearity that exists between genes and pathways using doubly ridge penalisation of entire pathways. We downloaded the executable file of PHARAOH from (http://statgen.snu.ac.kr/software/pharaoh/?page_id=11/) to perform the analysis. We matched each rare variant with the gene and then pathway based on their positions, during the analysis.

### 2.5 Multiple testing comparison

We performed three types of analysis, GWAS, gene and pathway level analyses. In the case of GWAS analysis, 577,605 SNPs were used. Likewise, when multiple comparison
tests are performed, there is a need to control for Family Wise Error Rate (FWER), the probability of at least one Type I error (false positive). To handle this issue, we calculated q-values using False Discovery Rate (FDR) correction, at a threshold q-value of 0.05, as our level of significance. This approach is applied to both gene and pathway analysis.

3 Results

3.1 Results of GWAS for common and less common variants

None of genetic variants was significant at $q$-value $\leq 0.05$, while 10 genetic variants were significant at $p$-value $\leq 10^{-5}$. Of the 10 variants, 7 variants were annotated to related genes (see Table 2). These seven variants were mostly located in the intron regions of genes and have never been introduced through clinical research. Thus, although some mutations regarded to be associated with susceptibility were identified, there was difficulty in explaining its biological association to genes and its relation to the susceptibility of COVID-19. Figure 1 shows the Manhattan plot of the GWAS results.

Table 2

<table>
<thead>
<tr>
<th>SNP</th>
<th>CHR</th>
<th>Base-pair position</th>
<th>Odds ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs9883738</td>
<td>3</td>
<td>1.77E+08</td>
<td>1.556</td>
<td>0.000002</td>
</tr>
<tr>
<td>rs12507905</td>
<td>4</td>
<td>25464527</td>
<td>1.203</td>
<td>0.000005</td>
</tr>
<tr>
<td>rs117461247</td>
<td>8</td>
<td>1.3E+08</td>
<td>1.453</td>
<td>0.00001</td>
</tr>
<tr>
<td>rs7357635</td>
<td>9</td>
<td>78774586</td>
<td>0.8606</td>
<td>0.000009</td>
</tr>
<tr>
<td>rs56321152</td>
<td>9</td>
<td>84200132</td>
<td>1.256</td>
<td>0.000007</td>
</tr>
<tr>
<td>rs73116579</td>
<td>12</td>
<td>59089418</td>
<td>1.613</td>
<td>0.000005</td>
</tr>
<tr>
<td>rs3751375</td>
<td>13</td>
<td>39606402</td>
<td>1.29</td>
<td>0.000001</td>
</tr>
</tbody>
</table>

Figure 1  
Manhattan plot of GWAS result. The red line is a significant $p$-value at Bonferroni correction level. The blue line shows $p$-value $1E-5$. SNPs with red circles are known to have related genes.
3.2 Results of gene-level and pathway analysis

3.2.1 Results for common and less common variants

In this analysis, none of genes and pathways was significant at a \( q \)-value \( \leq 0.05 \). We further examined pathways whose \( p \)-values were \( \leq 0.05 \) without FDR correction. As a result, we identified one such common pathway with both GSA-SNP2 and MAGMA. The pathway was Nod Like Receptors (NLRs) signalling pathway which belongs to one of the major innate immune systems (see Figure 2). And Figure 3 shows the QQ plots for each of the analysis including GWAS, gene-level analysis and pathway analysis for common and less common variants.

Figure 2 Scatter plot of \(-\log_{10} p\)-value from GSA-SNP2 and MAGMA. The \( X \)-axis is \(-\log_{10} p\)-values of GSA-SNP2. The \( Y \)-axis is \(-\log_{10} p\)-values of MAGMA. There are more points above the red line than below the red line, indicating that the \( p \)-value of MAGMA is often greater than \( p \)-value of GSA-SNP2. The dashed line represents the point where the \( p \)-value for each analysis is 0.05. The point marked with a red circle indicates the Nod Like Receptors signalling pathway.
Figure 3  QQ plots for each of the analysis including gene-level analysis and pathway analysis for rare variants

3.2.2 Results for rare variants

At gene-level association analysis, no genes were significant at $q$-value $\leq 0.05$. We focused on revealing significant pathways and seeking relationships between significant pathways and biological importance. Three pathway analysis methods for rare variants confirmed pathways that may be related to the susceptibility of COVID-19. As shown by Figure 4, the seven pathways were significant across all methods at $q$-value $\leq 0.05$ (see Table 3). The SKAT and SKAT-O analysis discovered 67 and 72 significant pathways respectively, while PHARAOH discovered 16 pathways, all significantly associated with the susceptibility of COVID-19, at the same $q$-value threshold (see Figure 4). Especially, only fewer pathways were significant in PHARAOH analysis compared to the other two analyses, but more than half of the pathways were likely to be associated with viral infection susceptibility, including COVID-19. Among the 8 significant pathways which were significant only in PHARAOH, 5 pathways including thyroid cancer and Ascorbate and Aldarate metabolic pathways were described to be related with COVID-19 or general viral infection process in previous studies. Details will be described in the discussion section.

PHARAOH which is a hierarchical structure model makes it possible to reveal the significance of genes and pathways simultaneously. Thus, we could confirm that 3 pathways of the 7 commonly significant pathways included 5 significant genes at a $p$-value of 0.05 (see Table 4, Figure 5). Some genes have been reported to be related with COVID-19 through previous studies (Johnson et al., 2020; Zauli et al., 2020). Figure 6 shows QQ plots of $p$-values which were obtained from each of the analysis including gene-level analysis and pathway analysis for rare variants.
Figure 4  Venn diagrams of results from pathway analysis using rare variants. We determined significant pathways whose $q$-value was less than or equal to 0.05. 7 pathways were commonly significant in three methods.

Table 3  List of significant pathways identified by rare variants analysis

<table>
<thead>
<tr>
<th>KEGG pathway name</th>
<th>PHARAOH</th>
<th>SKAT</th>
<th>SKAT-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG_OOCYTE_MEIOSIS</td>
<td>0.037</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>KEGG_JAK_STAT_SIGNALING_PATHWAY</td>
<td>0.037</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND_NEOLACTO_SERIES</td>
<td>0.037</td>
<td>0.005</td>
<td>0.008</td>
</tr>
<tr>
<td>KEGG_LONG_TERM_DEPRESSION</td>
<td>0.037</td>
<td>0.005</td>
<td>0.008</td>
</tr>
<tr>
<td>KEGG_ARGinine_AND_PROLINE_METABOLISM</td>
<td>0.046</td>
<td>0.008</td>
<td>0.001</td>
</tr>
<tr>
<td>KEGG_PYRUVATE_METABOLISM</td>
<td>0.037</td>
<td>0.027</td>
<td>0.001</td>
</tr>
<tr>
<td>KEGG_P53_SIGNALING_PATHWAY</td>
<td>0.037</td>
<td>0.011</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Table 4  List of genes included in significant pathways from PHARAOH analysis

<table>
<thead>
<tr>
<th>KEGG pathway name</th>
<th>Gene symbol</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG_GLYCOSPHINGOLIPID_BIOSYNTHESIS_LACTO_AND_NEOLACTO_SERIES</td>
<td>FUT9</td>
<td>0.034</td>
</tr>
<tr>
<td>KEGG_JAK_STAT_SIGNALING_PATHWAY</td>
<td>SOCS5</td>
<td>0.05</td>
</tr>
<tr>
<td>KEGG_P53_SIGNALING_PATHWAY</td>
<td>IGFBP3</td>
<td>0.05</td>
</tr>
<tr>
<td>KEGG_P53_SIGNALING_PATHWAY</td>
<td>BID</td>
<td>0.02</td>
</tr>
<tr>
<td>KEGG_P53_SIGNALING_PATHWAY</td>
<td>MDM2</td>
<td>0.05</td>
</tr>
</tbody>
</table>
**Figure 5** Network plot from PHARAOH analysis shows significance pathways ($q$-value ≤ 0.05) including at least one gene ($p$-value ≤ 0.05). Green nodes represent pathway and blue nodes represent genes. Genes are associated with phenotype (COVID-19 infection) through hierarchical structure.

**Figure 6** QQ plots for each of the analysis including GWAS, gene-level analysis and pathway analysis for common and less common variants.
4 Discussion

We performed GWAS, gene-level association and pathway analysis using COVID-19 tests result from UKB data, to identify associations between COVID-19 infection susceptibility and host genome. In GWAS analysis, there were no significant SNPs at \( q \)-value threshold of \( \leq 0.05 \).

We identified significant genes (using PHARAOH only) and pathways associated with the susceptibility of SARS-COV-2 using the available two types of variants, 1) common and less common and 2) rare variants. In common and less common variants analysis, there were no significant genes and pathways at \( q \)-value \( \leq 0.05 \). When we applied significance level without considering the multiple test correction (\( p \)-value \( \leq 0.05 \)) to the results from the two pathway analysis methods (MAGMA and GSA-SNP2), we confirmed that the NLR signalling pathway is significant in both methods.

Also, we identified seven pathways that were commonly significant at \( q \)-value \( \leq 0.05 \) from all pathway analysis methods, for rare variants. We investigated published studies to understand the biological roles played by these common significant pathways in viral infections (see Table 3). Among the seven pathways, two pathways have been reported to be related to viral infection in more than three papers (Hennighausen and Lee, 2020; Luo et al., 2021; Matsuyama et al., 2020). These studies reported the association of JAK-STAT signalling pathway with COVID-19. This pathway includes ACE2, which is used as the entry receptor by SARS-COV-2 virus. Meanwhile, we identified seven genes that were commonly significant in gene-level association analysis for rare variants after using the significance level without considering the multiple test correction (\( p \)-value \( \leq 0.05 \)), although there were no significant genes at \( q \)-value \( \leq 0.05 \). Among them, only B4GAT1 was included in glycosphingolipid biosynthesis, which was a significant pathway from pathway analysis.

Investigation of published studies reveals that NLR signalling pathway plays a major role in recognising virus in the innate immune system. Jacobs and Damania (2012) confirmed the role of NLR proteins in sensing different viral infection. In recent pandemic situations, NLRs have also been studied with consideration of the role of NLRs in COVID-19 infection. Li et al. (2020) suggested that the production of IFN which limits virus spread is related to Pattern Recognition Receptors (PRRs). And the PRRs mainly include TLRs, RLRs and NLRs. Among the main PRRs, TLRs signalling pathway was actively studied for its association with COVID-19 infection and were reported as a component related to the severity of COVID-19 by several studies. However, research about the association between NLRs and COVID-19 is still lacking. Our research suggests that common and less common variations included in the NLR signalling pathway, are associated with COVID-19 infections.

In addition, glycosphingolipid biosynthesis pathway has central roles in virus entry and therapeutic targets. Abu-Farha et al. (2020) investigated the role of lipid metabolism in COVID-19 infection and as targets of drugs. Lipids play a major role in viral infection as a structural foundation of cells and viral membranes (Lorizate and Krausslich, 2011). To enter the host cells and modify host cells to produce viral molecules, viruses attack lipid synthesis and signalling pathway (Murillo et al., 2015). During the early replication stage, SARS-COV makes lipid rafts which include sphingolipid, cholesterol, and protein-rich microdomains. This facilitates the fusion as well as the injection of viral genome into host cell. Not only can sphingolipids regulate important membrane properties, but their properties might be suitable as novel targets for therapeutic intervention.
Next, JAK-STAT pathway plays an important biological function in combating viruses from entering the body. When viruses enter host’s bodies, interferons are generated in response to anti-virus reactions. According to the study of Fleming (2016), generation of interferon stimulates the JAK-STAT signaling pathway, leading to expression of genes that can kill actual viruses. Thus, if the JAK-STAT pathway is modified, the ability to kill viruses within the body may be different from others. It shows that the host’s antiviral response-related pathway can be associated with the infectious nature of COVID-19.

Furthermore, we confirmed the association between JAK-STAT signalling pathway and the infection of COVID-19. Ziegler et al. (2020) reported that SARS-CoV-2 spike(s) protein binding angiotensin-converting enzyme 2 (ACE2) which is a cellular receptor for SARS-COV-2. According to Luo et al. (2021), ACE2 was overexpressed in Human Airway Epithelial (HAEs) cells after COVID-19 infection and ACE2 had a significant correlation with JAK-STAT signalling pathway. The results suggest that JAK-STAT signalling pathway downstream action is that of overexpression of ACE2.

Especially, more than half of the pathways which were significant in the PHARAOH analysis were associated with COVID-19 susceptibility according to the previous researches. One of them is thyroid cancer pathway. In previous studies, researchers suggested associations between thyroid disorders and viral infection disease including COVID-19. The thyroid hormone is a critical component to regulate the major metabolism and biochemical pathways. Kumari et al. (2020) reported the potential role of prevailing thyroid disorders in COVID-19 infection. Importantly, individual with abnormal thyroid hormone have over-expression of ACE2 protein which has a central role in host-cell entry of COVID-19. Also, they suggested that monitoring the thyroid hormone can help in understanding the pathogenesis of COVID-19. Other pathways, including Selenoamino acid metabolism, Ascorbate and Aldarate metabolism, steroid hormone biosynthesis and Butanoate metabolism were also reported to play a role in the early stage of the COVID-19 and/or other viral infection (Alexander et al., 2020; Chemudupati et al., 2020; Hoang et al., 2020; Mihalopoulos et al., 2020).

In conclusion, we have identified pathways and genes that are associated with COVID-19 infection. Some of these pathways or genes are under active investigation or already well known, while some of them such as NLR has not been fully understood about their role in infection of COVID-19. Therefore, our research should be validated in several ways in further analysis. First, we can further study the rare variants included in well-known pathways to understand the infection process of the COVID-19. Also, it is possible to use mouse models to study the role of genes and pathways at the early stage of COVID-19 infection. Munoz-Fontela et al. (2020) reviewed several mouse models that have similar characteristics to humans to study COVID-19. Using these mouse models, it will be able to regulate specific genes to measure the causality between genetic factors and infection of COVID-19. Kwok et al. (2020) suggested that understanding how human genetics influence infectious disease susceptibility offers the opportunity for new insights into pathogenesis, potential drug targets, risk stratification, response to therapy and vaccination.

As new infectious diseases continue to emerge, together with growing levels of antimicrobial resistance and increasing awareness of genetic diversity at both the individual and population level, these results and other similar studies will inspire a future with a more tailored application of the precision medicine approach that maximises effectiveness for a given person or population group.
Acknowledgements

This research has been conducted using data from UK Biobank, a major biomedical database. This research was funded by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare (No. HI16C2037), and the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation (NRF).

References


Analysis of COVID-19 genetic risk susceptibility


The UK Biobank (2014) UK Biobank Axiom® Array Content Summary, Biobank, U.


