EXPLORING THE GENETIC LANDSCAPE OF COVID-19 SUSCEPTIBILITY AMONG PATIENTS IN TÜRKIYE: A VARIANT DISCOVERY STUDY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

YAVUZHAN ÇAKIR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF MASTER OF SCIENCE
IN
BIOINFORMATICS

JULY 2024

Approval of the thesis:

EXPLORING THE GENETIC LANDSCAPE OF COVID-19 SUSCEPTIBILITY AMONG PATIENTS IN TÜRKIYE: A VARIANT DISCOVERY STUDY

Submitted by YAVUZHAN ÇAKIR in partial fulfillment of the requirements for the degree of **Master of Science in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Banu Günel
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim AYDIN SON
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yeşim AYDIN SON
Supervisor, **Health Informatics Dept., METU** _____

Examining Committee Members:

Assist. Prof. Dr. Aybar Can ACAR
Health Informatics Dept., METU _____

Assoc. Prof. Dr. Yeşim AYDIN SON
Health Informatics Dept., METU _____

Assist. Prof. Dr. İdil Yet
Bioinformatics Dept., Hacettepe University _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Yavuzhan Çakır

Signature :

ABSTRACT

EXPLORING THE GENETIC LANDSCAPE OF COVID-19 SUSCEPTIBILITY AMONG PATIENTS IN TÜRKIYE: A VARIANT DISCOVERY STUDY

Çakır, Yavuzhan

MSc., Department of Bioinformatics

Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son

July 2024, 58 pages

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has underscored the need to understand the genetic factors influencing disease susceptibility. This study adopts a data-driven approach to identify and analyze Single Nucleotide Variants (SNVs) associated with COVID-19 susceptibility in the Hacettepe University Hospital patient cohort. We systematically compiled and analyzed variants published in diverse scientific publications. We genotyped patients treated at the Hacettepe University Hospital (Ankara, Türkiye) using a multiplex approach with next-generation sequencing. The analysis included variant calling, linkage analysis, and statistical comparisons with non-Finnish European allele frequencies. Key findings suggest variants (rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917) exhibit different frequencies compared to the European population, suggesting potential genetic predispositions affecting COVID-19 susceptibility in the hospitalized population in Türkiye. Linkage disequilibrium analysis highlighted strong correlations between specific genetic loci, indicating inherited patterns. The study highlights significant genetic variations associated with COVID-19 susceptibility within a Turkish cohort, differing from European allele frequencies. These findings emphasize the importance of considering genetic diversity in public health strategies and enhance our understanding of the genetic factors that may influence disease susceptibility and severity. Further research with larger cohorts is recommended to validate these associations and explore their implications for disease management and prevention strategies.

Keywords: Pandemic, Covid-19, Susceptibility, Variants, SNV

ÖZ

TÜRKİYE'DEKİ HASTALARDA COVİD-19 DUYARLILIĞININ GENETİK DURUMUNUN ARAŞTIRILMASI: BİR VARYANT KEŞFİ ÇALIŞMASI

Çakır, Yavuzhan

Yüksek Lisans, Biyoenformatik Bölümü

Tez Yöneticisi: Assoc. Prof. Dr. Yeşim Aydın Son

Temmuz 2024, 58 sayfa

SARS-CoV-2 virüsünün neden olduğu COVID-19 pandemisi, hastalığa yatkınlığı etkileyen genetik faktörleri anlama ihtiyacının altını çizmiştir. Bu çalışma, Hacettepe Üniversitesi Hastanesi'ndeki hastaları kullanarak Türk toplumunda Tek Nükleotid Varyantları (SNV'ler) ile COVID-19 duyarlılığı arasındaki ilişkiyi aydınlatmayı amaçlamaktadır. Bu çalışma, COVID-19 duyarlılığı ile bağlantılı SNV'leri tanımlamak ve analiz etmek için veri odaklı bir yaklaşım benimsemektedir. Yeni nesil dizileme kullanarak, Türkiye'deki Hacettepe Üniversitesi Hastanesi'nde tedavi gören hastalara odaklanarak, çeşitli bilimsel yayınlardan SNV verilerini sistematik olarak derledik ve analiz ettik. Analiz, varyant arama, bağlantı analizi ve Finlandiya dışı Avrupa alel frekansları ile istatistiksel karşılaştırmaları içermektedir. Temel bulgular, varyantların (rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115 ve rs56106917) Avrupa popülasyonuna kıyasla farklı frekanslar sergilediğini ve Türkiye'de hastanede yatan popülasyonda hastalık duyarlılığını etkileyen potansiyel genetik yatkınlıkları gösterdiğini öne sürmektedir. Bağlantı dengesizliği analizi, belirli genetik lokuslar arasındaki güçlü korelasyonları vurgulayarak kalıtsal modellere işaret etmiştir. Çalışma, Avrupa'daki alel frekanslarından farklı olarak, Türk nüfusu içerisinde COVID-19 duyarlılığı ile ilişkili önemli genetik varyasyonları vurgulamaktadır. Bu bulgular, halk sağlığı stratejilerinde genetik çeşitliliğin dikkate alınmasının önemini vurgulamakta ve hastalığa yatkınlık ve şiddeti etkileyebilecek genetik faktörlere ilişkin anlayışımızı geliştirmektedir. Bu ilişkileri doğrulamak ve hastalık yönetimi ve önleme stratejileri üzerindeki etkilerini araştırmak için daha büyük kohortlarla daha fazla araştırma yapılması önerilmektedir.

Anahtar Sözcükler: Covid-19, pandemi, yatkınlık, genetik, SNV

To my cherished family of origin and the loving family I chose through my life.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Assoc. Prof. Dr. Yeşim Aydın Son, for her invaluable guidance, support, and encouragement throughout my research. Her expertise and insight have been instrumental in the completion of this thesis.

I deeply thank Cahit Burduroğlu for his extensive assistance with bioinformatics tools and wet lab work. His expertise and willingness to help at every step significantly impacted my research journey.

I acknowledge Ahmet Görkem Er, the project colleague, for our collaborative learning experience.

My heartfelt thanks to my friend Nisan Korkmaz Aladağ for her help with the content and format of this thesis.

To my beloved wife, Simay Hurşidi Çakır: You are the weave that holds my dreams, the anchor in my stormy seas, and the ground beneath my soaring ambitions. Your unwavering support and love have made this journey possible. Thank you for being my constant inspiration and guiding light.

To my parents, Ayşe and Çetin Çakır, your belief in me has been a constant source of strength.

I am equally grateful to my wife's family, Sara, Melek, and Huseyin Huşidi, for their kindness and support.

I would also like to thank Desia Clinical Research, my current workplace, for their support and understanding during the completion of this thesis.

I appreciate TÜBİTAK for awarding me a scholarship during my Master's program, which provided me with the resources to focus on my studies and research.

I am grateful to AdımODTÜ for their financial support, which helped fund a portion.

Special thanks to my supporting relatives Ürkiye, Murat, Bengü, Damla, and Ali Kaan ATEŞ. Your encouragement and belief in me have been invaluable.

Lastly, to my dog, Metis Hurşidi Çakır, thank you for being a constant source of joy and companionship. Your presence has been a great comfort during this journey.

Thank you all for your contributions and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BSC	Biosafety cabinet
Conc.	Concentration
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
HEPA	High-efficiency particulate air
KO	Knock Out
LD	Linkage disequilibrium
NCBI	National center for biotechnology information
NGS	Next generation sequencing
PCR	Polymerase chain reaction
SNV	Single Nucleotide Variant

CHAPTER 1

1. MOTIVATION

The COVID-19 pandemic, triggered by the SARS-CoV-2 virus, has posed unprecedented public health challenges globally. The genetic predisposition to infection and disease severity is a critical aspect of the disease, significantly influencing its management and control. This study employs a multiplexed next-generation sequencing approach, which investigates the association of Single Nucleotide Variants (SNVs) with COVID-19 susceptibility among patients treated at Hacettepe University Hospital in Turkey.

First, we identified SNVs that could predispose individuals to COVID-19 through a systemic literature review. Then, we primarily focus on a cohort of 120 patients admitted to intensive care and those receiving treatment in general wards. Through rigorous variant calling, linkage analysis, and frequency comparison with non-Finnish European populations, we identified several SNVs with significant variations in allele frequencies. Notably, SNVs such as rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917 exhibited distinct patterns, suggesting potential genetic factors influencing susceptibility to SARS-CoV-2.

Our results revealed that specific genetic loci, including those associated with the ACE2 and TMPRSS2 genes, known for their roles in viral entry into host cells, show significant differences in allele frequencies compared to European populations. For instance, rs17860115 displayed a higher frequency in the Turkish cohort, which may correlate with an increased susceptibility to COVID-19. Conversely, other variants were found more frequently in the European population, potentially indicating a protective role against the disease.

Linkage disequilibrium analysis further emphasized the inherited patterns of these SNVs, revealing strong correlations between specific loci on the same chromosome, which could influence the disease phenotype. Statistical analysis using the Z-test and power analysis with the bootstrap methods confirmed the significant differences in allele frequencies between the populations, underscoring the importance of considering ethnic and regional genetic backgrounds in disease risk assessments.

This study highlights the critical need for integrating genetic data into public health strategies and personalized medicine approaches. By understanding the genetic

factors influencing disease susceptibility, public health authorities can better tailor interventions and potentially mitigate the impact of outbreaks. The findings also pave the way for future research into the genetic mechanisms of COVID-19, which could lead to more effective therapeutic and preventive strategies.

The implications of this research are profound, suggesting that genetic diversity plays a significant role in disease dynamics, which can be leveraged to enhance disease prediction models and response strategies. Further research with larger and more diverse cohorts is recommended to validate these findings and explore the full potential of genetic markers in managing infectious diseases like COVID-19.

CHAPTER 2

2. INTRODUCTION

Advancements in molecular methodologies, particularly next-generation sequencing, have significantly influenced the fields of genetics and virology (Louten, 2023; Phalke et al., 2024). This study focuses on understanding how Single Nucleotide Variants (SNVs) affect susceptibility to COVID-19, a global health challenge posed by the SARS-CoV-2 virus. Without a preliminary hypothesis, the research adopts a data-driven approach to systematically compile and analyze SNVs from diverse scientific publications, exploring their potential impact on disease susceptibility.

This study aims to identify and analyze SNVs influencing susceptibility to COVID-19 among patients at Hacettepe Hospital in Türkiye. The research seeks to uncover the mechanisms of differential disease susceptibility among individuals by developing a comprehensive genetic profile based on this specific population. This targeted approach will enhance the understanding of genetic factors within the Turkish population and contribute to the broader scientific knowledge base, informing future genomic research and public health strategies against pandemic threats. This study is mainly aimed at tailoring public health responses and medical interventions to better meet the Turkish population's needs in the face of such global health challenges.

Before delving into this research, it is crucial to have a foundational understanding of COVID-19 and the concept of SNV. Grasping the nature and implications of this disease enhances comprehension of the study's findings and their significance in the broader context of global health challenges.

COVID-19 is a highly contagious disease caused by the SARS-COV2 virus that affects the upper respiratory tract. Its impact on the world has been socially and economically significant, particularly between 2020-2022. While no definitive measure exists to combat the disease, people worked diligently to find solutions and mitigate its impact (CDC, 2024; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020).

SARS-COV2 virus was first detected in Wuhan, China, in late 2019 and quickly spread worldwide, resulting in a global pandemic. SARS-CoV-2 is an RNA-based virus that belongs to the Coronaviridae family and is responsible for causing

COVID-19 disease. The virus has zoonotic origins which can be transmitted from animals to humans. These facts demonstrate the virus's ability to spread rapidly and its potential to cause harm, underscoring the importance of taking appropriate measures to prevent its transmission. The virus's outer surface has spike-shaped protein protrusions called Spike proteins. The Spike proteins are crucial in enabling the virus to attach to and penetrate human cells. COVID-19 is highly contagious and can be transmitted through respiratory droplets released when an infected person coughs, sneezes, or talks or by touching contaminated surfaces. Common symptoms of COVID-19 include fever, cough, shortness of breath, and loss of smell and taste. The severity of the disease caused by SARS-CoV-2 can vary, ranging from mild cases to severe lung infections and even death (CDC, 2024; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020; Spiteri et al., 2020).

The virus has a single-stranded RNA genetic structure consisting of approximately 30,000 nucleotides, which encodes proteins essential for the virus to replicate, enter cells, and evade the immune system. Different variants have emerged as the virus has mutated over time, resulting in genetic diversity. While some of these variants may spread more quickly, cause more severe disease, or develop resistance to existing vaccines and treatments, it is essential to note that the scientific community is actively working to address these challenges. Several significant variants of SARS-CoV-2 have been identified since the beginning of the pandemic, including Alpha, Beta, Gamma, Delta, and Omicron. The Delta and Omicron variants have demonstrated high transmissibility and resistance to some vaccines, but researchers are continuing to study and develop new strategies to combat these variants and protect public health. Implementing public health measures, vaccination campaigns, and treatment strategies is critical to control the virus's spread and reduce the disease's effects. These measures are of great importance in the fight against SARS-CoV-2 and COVID-19. It is essential to continue learning about the virus and its variants to develop effective strategies against the pandemic (CDC, 2024; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020; Mwendwa et al., 2024).

The incubation period of COVID-19 is approximately one week. Therefore, individuals may unknowingly transmit the disease to others during this time. The following measures were taken to prevent the spread of the virus and protect public health. Local and global efforts have been made to implement precautions. In Turkey, schools were closed temporarily on March 21, 2020, for three days, which became permanent. Curfews, public transportation bans, and restrictions on public gatherings followed this. These precautionary methods are implemented to control the spread of the disease (CDC, 2024; Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020).

During the early stages of the pandemic, SARS-COV2, the virus responsible for causing COVID-19 disease, was diagnosed and thoroughly examined for vaccine studies. Biontech, Johnson & Johnson, and Pfizer have developed vaccines targeting

the spike protein found on the membrane surrounding the SARS-COV2 virus and binds to the ACE2 receptor in human cells. These vaccines introduce mRNA molecules encoding the spike protein into the body to produce it within the cell and present it as an antigen. It should be noted that the evaluation of these vaccines is ongoing. Especially before the safety and performance tests of the vaccines were completed, as people were unvaccinated and the disease spread, many people around the world caught this disease, showing symptoms of varying severity. The most common symptoms of COVID-19 are upper respiratory tract symptoms such as cough, fever, loss of smell, and shortness of breath. People showed symptoms of different severities depending on many factors, such as age, smoking, and general body health. Some people survived the disease with only a minor cough problem, and some patients had to be connected to mechanical ventilators due to shortness of breath. In addition, in many countries around the world, including Türkiye, many economic, logistical, or social problems were also encountered (Aral & Bakır, 2022; Bostan et al., 2020; CDC, 2024; El-Maradny et al., 2024; Özen, 2024).

Due to the difference in the severity of the patient's symptoms, some patients were treated in intensive care units, and patients who tried to be examined or receive treatment in the hospital even though they were not treated in intensive care (Dasgupta et al., 2023; Kousathanas et al., 2022).

The variability in COVID-19 manifestations, even among individuals in the same household exposed to the same viral strains, highlights the complex interplay between genetic makeup and disease outcomes. This variability underscores the necessity to identify genetic markers that could predict an individual's susceptibility to viral infections. Understanding these genetic determinants is crucial for customizing public health responses, mainly when comprehensive epidemiological data is unavailable (Azzarà et al., 2022; Dasgupta et al., 2023; Kousathanas et al., 2022).

Utilizing SNV as susceptibility markers requires a detailed study of patient and healthy cohorts. SNVs, defined by single nucleotide alterations in the DNA sequence, contribute to genetic diversity within species. These genetic variations, typically occurring at intervals of 300 to 1000 base pairs along the genome, play crucial roles in gene function and are pivotal in understanding genetic disorders, drug responses, and susceptibility to environmental factors. SNVs are key markers for studying genetic diseases and provide insights into evolutionary biology and species adaptation. This research leverages next-generation sequencing to identify and analyze these variations, enhancing our understanding of their roles in health and disease (Brody, 2016; Børsting & Morling 2013; Gunderson 2007; Nelson et al. 2004).

SNVs are important genetic markers that can be used to understand genetic diseases, drug responses, and susceptibility to environmental factors. SNVs are crucial indicators of genetic diversity and evolutionary processes. They provide valuable insights into how species adapt and evolve. Additionally, specific SNVs have been

linked to an increased risk of developing specific diseases, such as cancer, heart disease, and diabetes. SNVs play a crucial role in understanding the genetic basis of diseases and personalized medicine (Brody, 2016; Børsting & Morling, 2013; Gunderson, 2007; Nelson et al., 2004).

Today, SNV biomarkers provide a comprehensive understanding of the genetic basis of diseases and enable personalized treatment options. So, SNV genotyping to analyze the presence of these SNVs when assessing an individual's susceptibility to these diseases is emerging as routine clinical tests. Genomic variations can be genotyped with various molecular techniques, including allele-specific PCR and DNA sequencing. In modern genetic research, next-generation sequencing (NGS) technologies are frequently used for SNV analysis, allowing the simultaneous detection of thousands of SNVs in large-scale genomic studies (Brody, 2016; Børsting & Morling, 2013; Gunderson, 2007; Nelson et al., 2004).

Some biological databases provide extensive data about the SNVs and other genetic variations. The dbSNV is one of the largest databases that collects and organizes this information. Here, information such as the locations of SNVs, allele frequencies, and even their possible health effects can be accessed. The 1000 Genomes Project, initiated to examine the genetic diversity of different human groups worldwide, provides a comprehensive source of genetic variation. Thanks to this project, valuable information such as SNV frequencies and distributions in various populations can be accessed. “rs” codes are a unique identifier assigned by dbSNV for each SNV and stand for “Reference SNV.” For example, a code such as "rs6265" identifies a specific SNV, and detailed information about that SNV can be obtained using this code. This can range from where the SNV is in the genome, what alleles it has, the genes it affects, and its potential health effects (Sherry et al., 2001).

The foundation of this study rests on an exhaustive literature search designed to identify SNVs associated with COVID-19. This search was structured to encompass a wide array of studies, ensuring a comprehensive compilation of SNVs linked to the disease's susceptibility and severity. By employing rigorous criteria for inclusion and a methodical approach to data extraction, this process has facilitated the identification of key genetic markers. These markers provide deeper insights into the genetic factors contributing to COVID-19 outcomes, laying the groundwork for the detailed analysis of implicated genes.

Several studies have highlighted the role of TMPRSS2 variants in COVID-19 susceptibility. For instance, rs2298661, identified near the TMPRSS2 and MX1 genes, has been linked to severe COVID-19 outcomes (Andolfo et al., 2021). Additionally, rs2298659 and other TMPRSS2 variants (rs34624090, rs35899679, rs4290734, rs463727) have been shown to modulate disease severity, potentially through androgen-responsive expression differences (Asselta et al., 2020; Barash et al., 2020).

The DPP4 gene, particularly the rs17574 variant of the DPP4 gene, has been associated with lower levels of sDPP4 and increased susceptibility to MERS-CoV, suggesting a possible mechanism for COVID-19 susceptibility (Alkharsah et al., 2021; Posadas-Sánchez, 2021).

The rs8178521 variant in the IL10RB gene has been significantly associated with long-COVID symptoms, indicating its potential role in prolonged COVID-19 effects (Angulo-Aguado et al., 2024).

Variants such as rs61299115, rs11088551, and rs4303794 have been identified in the regulatory regions of ACE2 and TMPRSS2, potentially affecting gene expression and influencing disease progression (Barash et al., 2020).

Studies like those by Kousathanas et al. (2022) have utilized whole-genome sequencing to uncover multiple genetic loci (e.g., rs1123573, rs12610495, rs17860115, rs2532300, rs56106917, rs61882275, rs9271609) associated with critical COVID-19, providing insights into the interferon signaling and leukocyte differentiation pathways.

Wooster et al. (2020) have linked specific ACE2 polymorphisms, including rs1548474, with varying levels of COVID-19 severity, offering potential markers for assessing hospitalization risks.

Studies have noted significant differences in the expression levels of key genes like ACE2, which do not show significant variations between males and females despite expectations based on its location on the X chromosome (Asselta et al., 2020).

Specific TMPRSS2 haplotypes and variants (e.g., rs463727, rs34624090) show different frequencies between populations, suggesting population-specific risks and responses to COVID-19 (Asselta et al., 2020; Iyer et al., 2020).

This review categorizes the literature based on genetic associations with either susceptibility or severity of COVID-19, highlighting the critical role of specific variants and the need for further studies across diverse populations to validate these findings. Each study contributes uniquely to our understanding of the genetic basis of COVID-19, offering potential targets for therapeutic and preventive strategies.

Following a comprehensive literature search to identify significant SNVs associated with COVID-19 susceptibility and severity, the focus shifts to the genes containing these variants. The following section aims to delve into these genes' biological functions and roles, exploring their contributions to the molecular mechanisms of SARS-CoV-2 infection and subsequent immune responses. By elucidating these genes' genetic architecture and expression patterns, insights into the pathways through which they influence disease dynamics and patient outcomes are sought. Understanding the regulatory roles, interactions with other cellular components, and impact on the body's defense mechanisms against COVID-19 is crucial for

developing targeted therapeutic strategies and enhancing predictive models of disease progression. Detailed insights into each gene identified in the SNV analysis will be presented, emphasizing their importance in the broader context of genomic medicine and public health.

ACE2 protein produced by the ACE2 gene is part of the angiotensin-converting enzyme family of dipeptidyl carboxydipeptidases and is closely related to the human angiotensin- I converting enzyme. This enzyme, which is secreted, converts angiotensin I to angiotensin 1-9 and angiotensin II to the vasodilator angiotensin 1-7. Known as ACE2, it is expressed in various human organs, and its specific expression in different organs and cells indicates a potential role in regulating cardiovascular and renal functions and fertility. Moreover, this protein serves as a functional receptor for the spike glycoprotein of several human coronaviruses, including HCoV-NL63, SARS-CoV, and SARS-CoV-2, the latter of which causes COVID-19. This gene has multiple splice variants; notably, the dACE2 (or MIRb-ACE2) variant is inducible by interferon (NCBI, 2024a).

BCL11A gene produces a C2H2 type zinc-finger protein, similar to the mouse Bcl11a/Evi9 protein. In mice, the corresponding gene frequently experiences retroviral integration at its location, which is linked to myeloid leukemia. It might contribute to leukemia development, partly through its interaction with BCL6. This gene is typically down-regulated during the differentiation of hematopoietic cells and could be involved in the development of lymphoma, as B-cell malignancy-associated translocations often disrupt its expression. Several transcript variants and different isoforms of this gene have been identified (NCBI, 2024b).

The DPP4 gene codes for the enzyme dipeptidyl peptidase 4, also known as adenosine deaminase complexing protein-2, and the T-cell activation antigen CD26. This enzyme is a type II transmembrane glycoprotein and a serine exopeptidase that removes X-proline dipeptides from the N-terminus of polypeptides. Dipeptidyl peptidase 4 plays a significant role in glucose and insulin metabolism and in immune system regulation. It has been identified as a functional receptor for the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), and protein modeling indicates it might interact similarly with SARS-CoV-2, the virus that causes COVID-19 (NCBI, 2024c).

The DPP9 gene produces a protein that belongs to the S9B family within clan SC of serine proteases. It exhibits post-proline dipeptidyl aminopeptidase activity, specifically cleaving Xaa-Pro dipeptides from the N-termini of proteins. While this protein's activity is similar to that of dipeptidyl peptidase 4 (DPP4), it is not membrane-bound. Dipeptidyl peptidases are generally involved in regulating the activity of their substrates and have been associated with various diseases, including type 2 diabetes, obesity, and cancer. This gene has several transcript variants, though they have not been fully characterized (NCBI, 2024d).

The ELF5 gene codes for a transcription factor belonging to the epithelium-specific ETS family. It plays a role in regulating the later stages of keratinocytes' terminal differentiation and influences various epithelium-specific genes in tissues with glandular epithelium, such as the salivary gland and prostate. This transcription factor has a low affinity for DNA, attributed to a negative regulatory domain at its amino terminus. It has also been identified as a tumor-suppressive transcription factor in breast cancer (NCBI, 2024e).

Despite the limited information available on the FBRSL1 gene, it has been identified to facilitate RNA binding activity (NCBI, 2024f).

HLA-DRB1 is a member of the HLA class II beta chain paralogs. This class II molecule is a heterodimer, comprising an alpha chain (DRA) and a beta chain (DRB), both membrane-bound. It plays a crucial role in the immune system by presenting peptides derived from extracellular proteins. These molecules are predominantly expressed in antigen-presenting cells. The beta chain weighs between 26-28 kDa and is structured into six exons: the first exon encodes the leader peptide; exons two and three code for the extracellular domains; exon four for the transmembrane domain; and exon five for the cytoplasmic tail. The beta chain carries all the polymorphisms that determine peptide binding specificities. Numerous DRB1 alleles have been identified, with some showing elevated frequencies linked to specific diseases or conditions, such as the DRB1*1302 allele's association with acute and chronic hepatitis B virus persistence. Additionally, this gene has several pseudogenes (NCBI, 2024g).

The protein produced by the IFNAR2 gene is a type I membrane protein that composes one of the two chains in the receptor for interferons alpha and beta. When activated, this receptor stimulates Janus protein kinases, which subsequently phosphorylate several proteins, including STAT1 and STAT2. This protein is part of the type II cytokine receptor family. Mutations in this gene have been linked to Immunodeficiency 45 (NCBI, 2024h).

The protein encoded by the IL10RB gene is a cytokine receptor family member. It is an essential accessory chain in the active interleukin 10 (IL-10) receptor complex. This protein and IL10RA need to be expressed together for the IL-10-induced signal transduction to occur. This gene and three other interferon receptor genes—IFAR2, IFNAR1, and IFNGR2—constitute a class II cytokine receptor gene cluster positioned on a small section of chromosome 21 (NCBI, 2024i).

The KANSL1 gene codes for a nuclear protein that forms part of two protein complexes associated with histone acetylation: the MLL1 and NSL1 complexes. The protein is involved in various cellular functions such as enhancer regulation, cell proliferation, and mitosis. Mutations in this gene have been linked to Koolen-de Vries Syndrome (NCBI, 2024j).

The TMPRSS2 gene codes for a protein that is a member of the serine protease family. This protein features several structural domains: a type II transmembrane domain, a receptor class A domain, a scavenger receptor cysteine-rich domain, and a protease domain. Serine proteases are implicated in numerous biological and disease processes. The expression of this gene is up-regulated by androgenic hormones in prostate cancer cells but down-regulated in androgen-independent prostate cancer tissues. The protease domain is believed to undergo autocleavage, resulting in its secretion into the cellular media. This protein also assists in the entry of viruses into host cells by cleaving and activating the viral envelope glycoproteins. Viruses such as Influenza and human coronaviruses HCoV-229E, MERS-CoV, SARS-CoV, and SARS-CoV-2 (COVID-19) utilize this protein for cell entry. Multiple isoforms of this protein have been identified due to alternative splicing (NCBI, 2024k).

CHAPTER 3

3. METHODS

3.1.Sample collection

Various methods are available for collecting DNA from volunteers in sequencing studies, such as blood samples, epithelial tissue samples, and saliva samples. In this study, epithelial tissue samples are obtained by scraping tissue from the inside of the cheek, or volunteers provide a saliva sample by spitting it into collection tubes. In cases where the volunteer cannot provide saliva consciously, suction methods are used to collect fluids. Collecting DNA from saliva samples is a non-invasive and user-friendly method frequently used in genotyping studies. Saliva collection kits typically include sterile collection tubes, instructions, and labeling materials. Participants are required to abstain from eating, drinking, chewing, or smoking for a minimum of 30 minutes before providing a saliva sample. The collection tube has a marked line indicating the required amount of saliva. To provide the sample, participants should remove the cap from the tube and spit directly into it, following the instructions. It is crucial to avoid touching the lips to the edge of the tube during this process to ensure the accuracy of the sample. The collection tube contains a solution. Once the participant provides a saliva sample, the solution inside the tube that preserves the DNA, keeping it stable, is mixed automatically or manually with the saliva. The sample is then promptly sent to the laboratory for processing or stored at the recommended temperature (Gudiseva et al., 2016; Rogers et al., 2007; van Oorschot et al., 2016).

For the COVID-19 disease, in this study, officially the first case of which was announced in Türkiye on March 21, 2020, public and private hospitals made great efforts to control the disease and its spread. One of these hospitals in Turkey was Hacettepe Hospital. Both intensive care patients and ward patients received treatment at Hacettepe Hospital. Samples were collected from 120 patients via NORGEN BIOTEK CORP. Saliva DNA kits. These collected samples include both intensive care patients and ward patients. The sample collection process involves collecting saliva by asking the patient to spit into a previously prepared container. However, this procedure was more difficult in intubated patients connected to a respiratory support unit, where suction of respiratory fluids was preferred.

3.2.DNA Isolation

DNA isolation separates and clears DNA from cells or virus particles in a sample. This process is a fundamental step for genetic analyses. The DNA isolation process usually begins with a cell-washing step. The cells or tissue from which DNA will be isolated in the cell washing and collection steps can be from various sources such as blood, tissue samples, and bacterial cultures. Cells are washed with a suitable buffer solution and collected by centrifugation. Next, it is necessary to break down the outer layers of the cell, including the cell and nuclear membranes, to reach the DNA. This is usually done using a lysis buffer containing detergent or enzymes (e.g., lysozyme or proteinase K). Detergents disrupt membranes by dissolving lipids; Enzymes break down proteins. After cell lysis, purifying the DNA from outside proteins and other molecules is necessary. This is done by precipitating the proteins and separating the DNA in the solution. One frequently used method is the addition of sodium acetate followed by precipitation of DNA with isopropanol or ethanol. The precipitated DNA is usually washed with ethanol. This process removes salts and other contaminants remaining in the DNA. The DNA is then extracted from ethanol, dried, and resuspended in a suitable buffer solution or water. DNA is usually brought to the purity and concentration required for specific applications in the final step. This may involve assessing the quantity and purity of DNA using spectrophotometry or agarose gel electrophoresis. Isolation protocols may vary for DNA obtained from different sources. For example, isolating DNA from blood samples may require additional steps to separate red blood cells from white blood cells, but there is no need for an extra process for DNA obtained from saliva (Garbieri et al., 2017).

DNA isolation in this study has been performed in the laboratory with a uniform isolation protocol as follows:

1. The preserved saliva sample was mixed by inversion and gentle shaking for a few seconds.
2. 0.5 mL of the preserved saliva sample was transferred to a 2 mL centrifuge tube (not provided).
3. 20 µL of Proteinase K was added (after vortexing Proteinase K) to the tube, mixed by vortexing for 10 seconds, and the sample was incubated at 55°C for 10 minutes.
4. 200 µL of Binding Buffer B was added to the sample. The mixture was vortexed for 10 seconds and then incubated at 55°C for 5 minutes.
5. An equal volume (total volume of preserved saliva, Proteinase K, and Binding Buffer) of room-temperature isopropanol was added to the sample. The sample was mixed gently by inversion ten times.
6. The mixture was centrifuged at room temperature for 4 minutes at a maximum speed of 20,000g (~14,000 RPM).

7. The supernatant was carefully removed and discarded, ensuring the DNA pellet was not disturbed. The tube was gently placed briefly upside down on a paper towel to remove residual isopropanol.
8. Twice, 500 µL of 70% ethanol was carefully added. The tube gently swirled and let stand at room temperature for 1 minute.
9. Twice, the sample was centrifuged at room temperature for 3 minutes at the maximum speed of 20,000g (~14,000 RPM), and then the ethanol was carefully removed and discarded without disturbing the pellet.
10. The open tube was placed upside down on a paper towel for 5 minutes to remove the excess 70% ethanol and air-dry the DNA pellet.
11. 50 µL of TE Buffer was added. The sample was vortexed for 30 seconds and incubated at 55°C for 5 minutes to rehydrate the DNA pellet, ensuring complete rehydration of the DNA before any subsequent steps.
12. The tube was centrifuged at 20,000 x g (~14,000 RPM) for 1 minute and 30 seconds to pellet any insoluble material.
13. The clear liquid was transferred into a clean tube, ensuring it did not disturb the pellet.
14. The purified DNA sample could be stored at 4°C for up to 2 months. For long-term storage, placing the samples at -20°C was recommended.

All these steps were conducted in a class II biosafety cabinet (BSC) in Düzen Laboratories since the saliva samples were from COVID-19 patients and COVID-19 patients are guaranteed to have SARS-COV-2 virus, an airborne virus.

Class II BSCs are designed to protect laboratory personnel, the environment, and the materials being worked on. They provide a barrier between the sample and the user, filtering the air with High-Efficiency Particulate Air (HEPA) filters, which capture potentially infectious particles. These cabinets have a vertical laminar airflow. The air is taken through a HEPA filter, providing a sterile work environment. It also has filtered inflow and downflow air, preventing any infectious particles from escaping from the cabinet. The HEPA filters can trap and retain particles as small as 0.3 microns with an efficiency of 99.97%, which is suitable for capturing airborne viruses like SARS-CoV-2. Class II BSCs are designed to provide a containment area for work with organisms that require Biosafety Level 1, 2, or 3 containments. For handling viruses like SARS-CoV-2, which typically require BSL-3 practices, these cabinets offer additional security when BSL-3 facilities are unavailable. These cabinets are versatile for various work involving microbiological samples and are widely used in virology, cell culture, and biotechnology. They comply with most safety standards required for handling pathogens.

After these steps, nanodrop measurements were conducted to observe the DNA products in each tube.

3.3.Primer designing

Primer design is of critical importance for the amplification (duplication) of the target DNA sequence in PCR (Polymerase Chain Reaction) and other molecular biological techniques. Primers are short oligonucleotide sequences specific to the target DNA and, when designed correctly, provide specific amplification of the targeted DNA region. For an effective primer design, the target region must first be determined. The region on the target DNA desired to be amplified must be clearly defined. This region may be a specific region of a gene or a mutation site. The rsID codes are used to identify and display mutation sites. When designing a primer, since it is necessary to design primers for the beginning and end of the region to be amplified on the DNA, a pair of primers should be designed to cover the region in between. Primers are generally 18-25 base pairs long. This length balances sufficient specificity and effective binding. The melting temperature of the primers, the temperature required to transition to the single-stranded state, should ideally be between 50-60°C.

Tm (Melting Temperature) values of both primers should be close. Since the number of hydrogen bonds effectively determines the Tm value, it is crucial to know how many base pairs consist of adenine, thymine, or guanine and cytosine. GC content in primers should be between 40-60% if possible. Since G and C bases form stronger bonds compared to A and T, this balanced ratio ensures that the primer binds tightly to the target DNA. The presence of a G or C base at the 3' end of the primer is called a "GC clamp" and helps strengthen binding during amplification. Primer sequences can interact within themselves (secondary structures) or with each other (primer-dimers) in undesirable ways. This may reduce amplification efficiency. The formation of such structures should be minimized during primer design. In addition, if there are repetitive regions in the DNA, choosing before or after the repetitive regions may be beneficial for primer design, as primers can also attach to these regions. Various computer software and databases are used for primer design. These tools are valuable for checking the specificity of primers, calculating the risk of secondary structure formation, and identifying possible target sites within the genome of the targeted organism. In silico, analyses are performed to verify the specificity of the designed primers. This is important to ensure that primers only bind to the targeted region and not mistakenly bind to similar sequences (Chuang et al., 2013; Dieffenbach & Lowe, 1993).

In this study, PyroMark Q24 software is used during the primer design. The designed primers and parameters of designing are shown in Appendix E. In primer designing, the following parameters were used:

- PCR Primer Settings:
    - Min Primer Length 18
    - Max Primer Length 24
    - Optimal Amplicon Length From 80

    - Optimal Amplicon Length To 250
    - Max Amplicon Length 600
    - Allow Primer Over SNP No
    - Melting Temp Algorithm Nearest Neighbor
    - Primer Concentration [µM] 0.2
    - Min Melting Temperature [°C] 56.0
    - Max Melting Temperature [°C] 86.0
    - Max Allowed Tm Difference [°C] 10.0
    - Max GC Difference [%] 30
- Sequencing Primer Settings:
    - Min Primer Length 15
    - Max Primer Length 20
    - Min Distance From Target 0
    - Max Distance From Target 3
    - Allow Primer Over SNP No
    - Generate Forward Primers Yes
    - Generate Reverse Primers Yes
    - Melting Temp Algorithm Nearest Neighbor
    - Min Melting Temperature [°C] 35.0
    - Max Melting Temperature [°C] 65.0

3.4.PCR

Polymerase Chain Reaction (PCR) is a molecular biological technique that amplifies a specific DNA sequence into millions of copies, which can amplify even small amounts of DNA. PCR was developed by Kary Mullis in the 1980s and revolutionized modern biology. The PCR process uses a reaction mixture containing specially designed short pieces of DNA (primers), a DNA polymerase enzyme, four types of nucleotides (dNTPs), and target DNA. The PCR process occurs in cycles with denaturation, annealing, and elongation steps. In the denaturation step, the temperature is approximately 94-98°C, and the aim is to separate the double-stranded DNA into two single-stranded strands by heating the reaction mixture. In the annealing step, the temperature generally varies between 50-65°C, and by lowering the temperature, specific binding of the added primers to the single-stranded strands of the target DNA is ensured. This binding is based on the primers designed to complement the target DNA sequence. In the elongation step, the temperature is generally kept at 72°C, and the purpose of this step is to synthesize new DNA strands using free dNTPs, depending on the DNA polymerase enzyme (usually TAQ polymerase) primers. This step involves creating copies of the target DNA sequence. These three steps are repeated for 25-35 cycles. At each cycle's end, the target DNA sequence amount theoretically doubles. Thus, after an average of 30 cycles, the amount of DNA sequence theoretically increases to 2 to the 30th power, approximately 1 billion copies. Once the PCR process is completed, the presence and size of the amplified DNA products can be checked by agarose gel

electrophoresis. This technique uses electric current to separate DNA fragments according to their size, visualized with a DNA stain (Clark, 2019; Khehra et al., 2023; Wages, 2005).

In this study, according to the nanodrop results, some of the PCR products have been diluted before PCR experiments. If the nucleic acid (ng/uL) amount exceeds 25, the sample is diluted to obtain the desired 25 level. No concentration protocols have been conducted for samples with less nucleic acid than the desired levels.

With the isolated DNA samples, PCR was performed on each patient and each region to be investigated. For 120 patients and 19 regions of interest, 2280 PCR experiments were conducted. The following protocol was followed for each PCR to obtain PCR products:

- 12.5 µL of PyroMark PCR Master Mix was added.
- 2.5 µL of CoralLoad Concentrate was included.
- µL of Q solution was mixed in, bringing the total volume to 20 µL.
- For each tube, 1 µL of primer was added, making the total volume 21 µL.
- µL of template DNA was added (the concentration of the gDNA was specified), resulting in a final volume of 25 µL.

1. An initial PCR activation step was performed for 15 minutes at 95°C, where the HotStartTaq DNA Polymerase was activated.
2. The 3-step cycling process involved:
    a. Denaturation for 30 seconds at 94°C.
    b. Annealing for 30 seconds at 60°C for genomic DNA
    c. Extension for 30 seconds at 72°C.
3. The number of cycles conducted was 45.
4. A final extension step was performed for 10 minutes at 72°C.
5. After the final step, products were stored at 4°C for up to 1 week.

3.5.Gel electrophoresis

Agarose gel electrophoresis is a laboratory technique separating macromolecules such as DNA, RNA, or proteins according to size. This method is often used in molecular biology and genetics research to control the size of PCR products, separate DNA fragments, or analyze restriction enzyme digests. It consists of gel preparation, sample preparation, sample loading, electrophoresis, and observation steps. In the gel preparation step, agarose powder is mixed with TAE (Tris-acetate-EDTA) or TBE (Tris-borate-EDTA) buffer solution and heated. After complete dissolution, the solution is allowed to cool slightly. In this step, dyes such as ethidium bromide can be added to the solution. The slightly cooled but still liquid solution is poured into the gel mold. A comb is also placed in the gel container to create spaces for sample loading. DNA samples are prepared by mixing them with a loading buffer. Loading buffer allows the samples to sink in the gel and may contain a staining agent so that

DNA bands can be observed later. The prepared samples are carefully loaded into the gel wells. The gel is placed in the electrophoresis tank and filled with a suitable buffer solution. When the electrical source is connected to both ends of the tank, with one polar end placed, the negatively charged DNA molecules move toward the positive electrode (anode). DNA fragments move at different speeds depending on their size; Small pieces move faster, and larger pieces move slower. After electrophoresis, the gel is examined under UV light, and DNA fragments can be seen as bands. During the loading phase, a control ladder is loaded into one of the wells. Since this control ladder consists of pre-designed DNA sequences of specific sizes, an idea can be obtained about the approximate sizes of the bands in the resulting image. It can be used to check the size of the DNA fragments obtained as a result of PCR and to verify whether the amplification was successful (Drabik & Bodzoń-Kułakowska, 2016; Lee et al., 2012).

Gel electrophoresis tests were performed on several samples to observe whether the PCR experiments were successful. The protocol used for gel electrophoresis is as follows:

1. 2 g of agarose was weighed on a precision scale and transferred to an Erlenmeyer flask.
2. 100 mL of 1X TAE buffer was removed and mixed with the agarose.
3. The mixture was gently shaken and then placed in a microwave oven, operated at a high setting for 2.30 minutes.
4. After microwaving, the flask was taken out to ensure the agarose was completely melted.
5. The solution was then gently shaken again and allowed to stand at room temperature for 5 minutes.
6. 4 µL of Ethidium bromide was added within a fume hood and the solution was shaken gently.
7. The prepared gel was slowly poured into the cassette tank.
8. Any bubbles present were removed with a clean pipette tip, and the gel was allowed to solidify.
9. Once the gel had solidified, the comb was removed from the gel.
10. The gel was placed in the electrophoresis tank, and a 1X TAE Buffer was added to fill the tank's volume.
11. PCR products were loaded into the wells, with the ladder being loaded first from left to right.
12. The electrodes were connected to the power supply, and the gel was run at 150V for 40 minutes.

13. Upon completion of the electrophoresis, the gel was imaged using the Biorad Gel Doc EZ instrument.

3.6.Sequencing

Sequencing determines the nucleotide sequences of DNA or RNA molecules at one point within a specific or entire range. High-throughput sequencing technologies became one of the milestones for its effects on molecular biology and clinical sciences, such as the diagnosis of genetic diseases or variance analysis (Brown, 2022).

The history of sequencing begins with the introduction of the chain termination or "Sanger" method developed by Frederick Sanger in 1977. This method revolutionized the determination of DNA sequences and earned Sanger the Nobel Prize in Chemistry in 1980. The Sanger sequencing method has been the basis of DNA sequencing studies for several decades. During this period, with the development of automatic sequencing devices, sequencing became a faster and less labor-intensive process. The Human Genome Project (1990-2003) was one of the large-scale sequencing projects that challenged the capabilities of the Sanger method. The completion of the project was an important milestone in understanding genetic science. Following the completion of the Human Genome Project, the need for faster and more economical sequencing methods has led to the development of "next-generation sequencing" (NGS) or second-generation sequencing technologies (Brown, 2022; Heather & Chain, 2016).

Illumina sequencing is one of the most widely used NGS technologies designed to meet high-throughput DNA sequencing needs. Also known as second-generation sequencing, this method allows millions of DNA fragments to be sequenced quickly and cost-effectively simultaneously. DNA is divided into short pieces in the first step of Illumina Sequencing. Adapters are added to each DNA fragment. DNA fragments are fixed to a surface and amplified via PCR, creating multiple copies of each fragment. Each amplified DNA fragment is read individually during the cyclic sequencing process. In this process, the sequence of each nucleotide is determined using four different colored fluorescent markers. The resulting raw sequence data is analyzed using computer algorithms, and the complete sequence of the targeted DNA is obtained (Hughes et al., 2013).

Multiplex sequencing is a technique that allows multiple regions of DNA to be sequenced simultaneously. This method saves time and cost, especially in large-scale genetic analyses. As part of next-generation sequencing (NGS) technologies, multiplex sequencing allows multiple samples or target regions to be analyzed in parallel. This process finds applications in many fields, such as detecting genetic diseases, population genetic studies, and microbiome analyses. Thanks to the multiplex sequencing method applied when PCR products amplified for more than one region of the same organism are found in a single solution. It is aimed to

complete the entire sequencing at once instead of repeating the process as many times as the number of regions to be sequenced (Arredondo-Alonso et al., 2021; Beck 1993; Church & Kieffer-Higgins, 1988).

For sequencing, Genera laboratories conducted Illumina multiplex sequencing protocols. This method was followed since outsourcing the sequencing for multiplex sequencing was both time- and cost-efficient. For sequencing, PCR products of individual patients were combined in tubes and sent to Genera. In other words, for each patient, one tube that brings all SNV PCRs together has been filled. The sequencing results were shared in fastq.gz format as read by the Genera laboratories. In this study, all targeted regions of the patient have been sequenced in one reading and separated analysis despite the common standard that a single gene is targeted at a time and 96 wells are parallelized.

3.7.Variant Calling

Variant calling is a process used in genomics to identify variants from sequence data—this means finding the differences like SNVs, insertions, deletions, and large structural changes in a DNA sequence compared to a reference genome. This process is crucial for understanding genetic variations that may influence traits, contribute to disease, or offer insights into evolutionary biology. For variant calling, the reads should be aligned to a reference genome, a standard sequence representing the idealized sequence of a species. This alignment shows where the reads match the reference genome and where there are differences. Once alignment is complete, the next step is identifying where the DNA sequence differs from the reference sequence. This involves analyzing the aligned reads to detect mismatches (SNVs), insertions, deletions, and other structural variants. To identify potential variants, software tools compare the sequenced reads against the reference genome. Not all identified variants are actual genetic differences. Some may be errors introduced during sequencing or alignment. Variant calling tools apply filters to reduce false positives. The identified variants are then annotated to provide information on their potential biological impact, such as whether they occur in coding or regulatory regions and if they are known to be associated with diseases. Variant calling is a complex but essential process in genomics, enabling researchers to uncover the genetic basis of diseases, understand genetic diversity, and explore evolutionary relationships. The choice of tools and specific workflow can vary depending on the type of sequencing data and the research objectives (Koboldt, 2020; Zverinova & Guryev, 2021).

In this study, the variant calling steps were as follows:

1) fastq.gz files were inspected for quality and contamination information.
2) Reads were trimmed if required quality scores or contaminations were not achieved.

    1. Raw reads with contamination were trimmed with the Cutadapt version 2.6 tool.
3) To generate reference indexes and alignment, the bwa version bwa-0.7.17-r1188 tool has been utilized to align the reads with the reference human genome (Hg38).
4) To compress and sort sam files samtools version 1.19.2 has been utilized. In 3 steps, BAI files have been generated:
    1. BAM generation.
    2. BAM compression and sorting.
    3. BAI generation.
5) To generate VCF files, freeBayes tool was used.
6) After creating the VCF files, read depths from initial fastq.gz files were inspected, and locations with read depths lower than ten were ignored for the rest of the study since the threshold of 10 determined bad results.
7) Locations of interested regions were inspected utilizing Python programming and the pandas library and information of variants were saved to a csv file.

Codes and exact tool data for these steps can be found in the appendices.

3.8.Linkage analysis

"Linkage," or connection between variants, is an important concept that enables understanding how genetic variations are related. This term specifically refers to how often two or more genetic markers (e.g., SNVs) coexist. This association plays a critical role in understanding how genetic material is passed from parents to offspring and helps explain how genetic diseases, traits, or resistances are inherited. Linkage disequilibrium (LD) is the nonrandom coexistence of two or more loci with each other. If two loci are very close to each other, recombination (mixing genetic material during the transfer of genetic material from parents to offspring) may not separate them, and these loci are often inherited. LD is used to understand how often genetic variants coexist in each population and how these variants spread (Reich et al., 2001).

In this study, ldlink's LDmatrix tool was used to inspect the linkages of SNVs in the same chromosomes. Only the SNVs on the same chromosomes have been inspected because there would be no linkage between SNVs on different chromosomes.

3.9.Frequency analysis

For frequency analysis, as mentioned in the Variant Calling title, information from allele regions with read depth scores above ten is recorded in a csv file as homozygous wild-type, homozygous variant, or heterozygous regions. Then, this is done for each rs code (for the relevant gene region). Variant frequencies of alleles were examined. In this frequency analysis, the formula of the number of heterozygous patients plus the number of patients with the homozygous variant

multiplied by two divided by n was used, with the total number of alleles read at one point being n. Thus, frequency information was obtained for each RS code using the information obtained from the readable regions.

3.10. Statistical analysis

In this study, we employed various statistical methods to compare allele frequencies between the Turkish cohort and the non-Finnish European population, assess the significance of these differences, and validate our findings through robust techniques. For instance, z-test was used to compare with the European population by using the frequencies of the variants whose frequencies were determined and the number of alleles in which these variants (or patients with wild-type in these locations) were seen.

Minor Allele Frequency (MAF) is a measure of the frequency at which the second most common allele occurs in a given population. It is a key metric in genetic studies as it helps identify variants that may contribute to disease susceptibility. In this study, MAF was calculated for each Single Nucleotide Variant (SNV) in both the Turkish cohort and the non-Finnish European population. MAF values have been calculated by the following formula:

$$MAF = \frac{Number\ of\ minor\ alleles}{Total\ number\ of\ alleles}$$

Given the sufficient number of participants in our study, we were able to detect target variants and make direct MAF comparisons without the need for extensive GWAS methodologies. The primary statistical methods used in our analysis include the Z-test for comparing allele frequencies and a bootstrap method for validation.

We also designed a bootstrapping-based analysis to test the z-test results using a bootstrap study. Bootstrapping is a resampling technique used to estimate the distribution of a statistic by repeatedly sampling with replacement from the data. This method is particularly useful for validating the results obtained from other statistical tests. In this study, the bootstrap method was used to generate confidence intervals for allele frequencies and validate the Z-test results. The following steps were followed to compare the simulated data of each gene location with the European population within the framework of Bootstrap:

1. To create n number of simulated alleles, the NumPy library was used in Python programming, and the probability of not being wild type, p-value, and frequency in sequenced patients was used.
2. A total of k cohorts was created by repeating n alleles k times, and the allele frequency in each of these k cohorts was recorded.
3. 95% and 99% confidence intervals were determined from the recorded frequencies.

4. It was observed whether the allele frequency seen in the European population fell within these intervals.

CHAPTER 4

4. RESULTS

4.1.Nanodrop Quality Control

From 120 patients, saliva samples have been collected and DNA is isolated in class II BSCs with a uniform isolation protocol. These samples were collected in Hacettepe Unviersity Hospital and DNA is isolated in Düzen Laboratories. After the host DNA isolation, isolated DNA products have been relocated to CanSyl lab in Middle East Technical University. For the products that are relocated to CanSyl labs, Nanodrop experiments were conducted to observe the DNA product amounts. Nanodrop results were mostly satisfying regarding the nucleic acid amount in DNA isolate solutions. Sixty-one DNA isolates had more than 25 ng/uL nucleic acid in the tube. Five solutions had between 20 and 25 ng/uL, and the rest were less than 20 ng/uL. Although less than 20 ng/uL is not optimum for PCR experiments, this does not mean that PCRs will not work intendedly or that there will be errors. Instead, this means that we should be aware of the risks of PCRs resulting in insufficient or no amplifications. A detailed table of results for nanodrops can be found in the appendices.

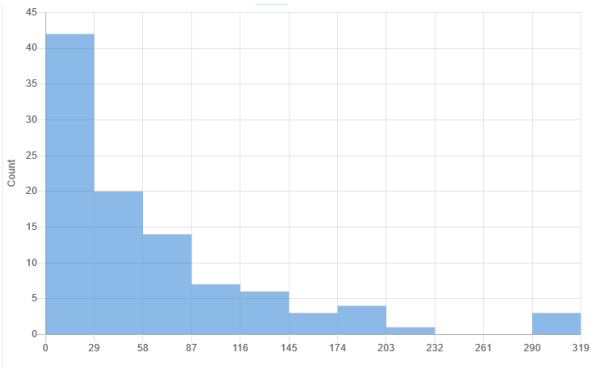

Figure 1 Histogram of nanodrop results in Nucleic Acid amount (ng/uL).

In figure 2, nanodrop results that are below 30 have been shown. Most of the DNA isolates with concentrations falling into the 'unsure' category, defined as being below the optimal threshold for reliable PCR amplification, are found between 5 and 7.5 ng/uL. This concentration range is critical, as it indicates a significant uncertainty in

the potential success of PCR amplification. While it is not necessarily prohibitive for PCR, it does increase the risk of suboptimal results, such as weak or no amplification. The distribution of these unsure data points is particularly important for assessing the need for further quality control measures such as gel electrophoresis confirmation after PCR.

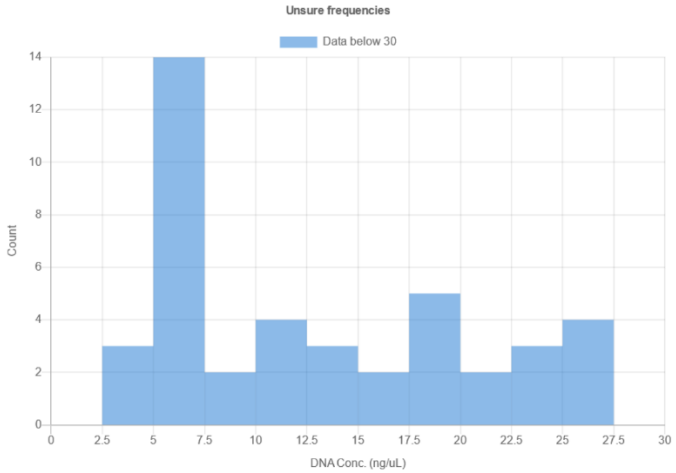

Figure 2 DNA concentration histogram from nanodrop results that are below 30 ng/uL

4.2.PCR amplification of target variant regions

The successful amplification of the target variant regions using PCR is critical for the subsequent sequencing and variant analysis. In this study, we performed PCR on DNA samples from 120 patients to amplify 19 regions of interest. Each PCR reaction was designed to target specific Single Nucleotide Variants (SNVs) associated with COVID-19 susceptibility. To ensure robust amplification, the PCR conditions were carefully optimized according to the steps mentioned in methods. PCR amplifications have been controlled and confirmed by electrophoresis experiments. In 29 patients' samples, for rs1123573, rs2532300, rs9271609, and rs17860115 either there was not enough material for PCR experiments or PCR experiments went faulty such that human eye can identify errors. Therefore, these 29 patients were not included in analyses regarding these SNVs' results.

The sequencing of the PCR-amplified target variant regions yielded predominantly satisfactory results. Multiplex sequencing was conducted to ensure comprehensive coverage and accurate identification of SNVs in the patient cohort. The majority of samples exhibited excellent read depths, which facilitated reliable variant calling and subsequent analyses. However, a few samples presented errors primarily due to insufficient read depths. To address this, a stringent quality control process was implemented. Sequencing data were carefully reviewed, and read depths were meticulously examined. Regions with read depths below the threshold of ten were identified as unreliable. Consequently, these low-quality reads were omitted from

further analysis to maintain the integrity and accuracy of the dataset. The read depth analysis ensured that only high-confidence variant calls were included in the final results. This rigorous filtering process is summarized in Appendix D, which provides detailed read depth information for each sample and target region. The table highlights the read depth distribution, identifying any samples with insufficient coverage that were excluded from the final dataset. Overall, the sequencing process demonstrated robust performance, with the majority of target regions being accurately sequenced and analyzed. The quality control measures effectively mitigated the impact of sequencing errors, ensuring the reliability of the genetic insights derived from this study.

4.3.Electrophoresis results

All electrophoresis experiments gave the desired results. Thus, it was seen that all PCR experiments worked properly, and a suitable working environment was created to continue with the sequencing process. Multiplex Sequencing and Variant Calling

4.4.MAF Calculations and Comparisons

For each rs code and corresponding patient, a comprehensive analysis was conducted to determine whether the individual exhibited wild-type alleles or polymorphisms within that specific genomic region. Concerning the single nucleotide variants (SNVs) examined in this study, the patients were categorized based on the presence of homozygous wild-type alleles, heterozygous alleles, and homozygous variant alleles, with their respective frequencies meticulously documented. Subsequently, these individual frequencies were aggregated to ascertain the total allele frequency for the cohort for each SNV. Moreover, to provide further insight into the data, the number of patients contributing to these frequencies was also recorded, specifically indicating the success rate of the readings, and this information was systematically included in the accompanying table. It should be noted that the frequencies displayed in the table below represent the number of patients, not the total number of alleles examined; thus, the actual number of alleles is twice the number of counts presented.

Table 1 Frequency information.

Region	Patient Count	Wild type homozygous patient frequency	Heterozygous patient frequency	Variant homozygous patient frequency	Variant allele frequency
rs11088551	120	0.33	0.51	0.15	0.404167
rs1123573	97	0.42	0.45	0.11	0.340206
rs12610495	117	0.63	0.33	0.03	0.192308
rs1548474	110	0.61	0.17	0.21	0.295455
rs17574	114	0.49	0.43	0.07	0.285088

Region	Patient Count	Wild type homozygous patient frequency	Heterozygous patient frequency	Variant homozygous patient frequency	Variant allele frequency
rs17860115	33	0.27	0.12	0.58	0.636364
rs2298659	118	0.72	0.25	0.02	0.144068
rs2298661	117	0.72	0.26	0.02	0.145299
rs35899679	116	0.4	0.43	0.16	0.37931
rs4290734	97	0.96	0	0.03	0.030928
rs4303794	120	0.33	0.51	0.15	0.404167
rs463727	117	0.36	0.41	0.22	0.42735
rs61882275	112	0.31	0.53	0.15	0.415179
rs8178521	120	0.49	0.41	0.09	0.295833
rs9271609	76	0.61	0.38	0	0.190789
rs2532300	106	1	0	0	0
rs34624090	116	1	0	0	0
rs61299115	120	1	0	0	0
rs56106917	84	1	0	0	0

The table below shows the z-test comparisons of SNVs' and European (non-Finnish) allele frequencies.

Table 2 Frequency comparisons of Hacettepe patients and non-Finnish European population.

SNV	Allele count	Variant allele frequency	Non-Finnish European frequency	P value of z test
rs11088551	240	0.40	0.42	0.69
rs1123573	194	0.34	0.39	0.36
rs12610495	234	0.19	0.27	0.06
rs1548474	220	0.30	0.29	0.90
rs17574	228	0.29	0.34	0.22
rs17860115	66	0.64	0.32	**0.00***
rs2298659	236	0.14	0.23	**0.03***
rs2298661	234	0.15	0.22	**0.04***
rs35899679	232	0.38	0.47	0.06
rs4290734	194	0.03	0.49	**0.00***
rs4303794	240	0.40	0.42	0.69
rs463727	234	0.43	0.46	0.48
rs61882275	224	0.42	0.37	0.37
rs8178521	240	0.30	0.27	0.49
rs9271609	152	0.19	0.30	**0.03***
rs2532300	212	0.00	0.22	**0.00***
rs34624090	232	0.00	0.45	**0.00***
rs61299115	240	0.00	0.42	**0.00***
rs56106917	168	0.00	0.49	**0.00***

According to the p values obtained from individual Z-tests conducted with the frequencies and the sample size of Hacettepe patients' alleles, the following SNVs show statistically significant differences with p-value < 0.05: rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917.

While rs17860115 shows a higher variant rate in the Hacettepe cohort, other SNVs show lower variant rates.

4.5.Bootstrap algorithm results

As mentioned, a bootstrap-based analysis is developed and applied with the existing variant frequency. With variant frequencies obtained from the frequency analyses, 1000 cohorts of 100 alleles were created, and 95% confidence intervals of frequency means of cohorts were compared with non-Finnish European allele frequencies. The following table shows if the non-Finnish European allele frequencies are within the limits of 95% confidence intervals of simulated data.

Table 3 Bootstrap results.

index	95% CI Lower bound	95% CI Upper bound	non-Finnish European allele frequency	Significant Difference
rs11088551	0.30975	0.5	0.422	No
rs1123573	0.25	0.43	0.385	No
rs12610495	0.12	0.27	0.269	No
rs1548474	0.21	0.38	0.29	No
rs17574	0.2	0.37	0.339	No
rs17860115	0.54	0.73	0.324	Yes
rs2298659	0.08	0.21	0.231	Yes
rs2298661	0.08	0.22	0.223	Yes
rs35899679	0.29	0.47	0.466	No
rs4290734	0	0.07	0.489	Yes
rs4303794	0.31	0.5	0.422	No
rs463727	0.34	0.52	0.46	No
rs61882275	0.33	0.51	0.374	No
rs8178521	0.22	0.39	0.268	No
rs9271609	0.12	0.27	0.304	Yes
rs2532300	0	0	0.218	Yes
rs34624090	0	0	0.449	Yes
rs61299115	0	0	0.421	Yes
rs56106917	0	0	0.493	Yes

As can be seen in the table above, bootstrap confidence intervals of SNVs rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917 does not encompass the exact non-Finnish European frequency.

4.6.Linkage results

SNVs rs17860115, rs8178521, rs463727, rs2298659, rs2298661, rs4290734, rs35899679, rs4303794, rs11088551, and rs61299115 are on chromosome 21. The following table is a matrix from ldlink (https://ldlink.nih.gov/?tab=ldmatrix) that shows the R^2 values of linkages, which is the possibility of carrying variants together. The closer the number to 1 means the closer two SNVs are also means the probability.

Table 4 Ldlink matrix results.

RS codes	**rs17860115**	**rs8178521**	**rs463727**	**rs2298659**	**rs2298661**	**rs4290734**	**rs35899679**	**rs4303794**	**rs11088551**	**rs61299115**
rs17860115	1.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rs8178521	0.01	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
rs463727	0.00	0.00	1.00	0.20	0.08	0.55	0.44	0.02	0.02	0.02
rs2298659	0.00	0.00	0.20	1.00	0.60	0.22	0.17	0.00	0.00	0.00
rs2298661	0.00	0.00	0.08	0.60	1.00	0.23	0.18	0.00	0.00	0.00
rs4290734	0.00	0.00	0.55	0.22	0.23	1.00	0.81	0.00	0.00	0.00
rs35899679	0.00	0.00	0.44	0.17	0.18	0.81	1.00	0.00	0.00	0.00
rs4303794	0.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	1.00	1.00
rs11088551	0.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	1.00	1.00
rs61299115	0.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	1.00	1.00

As shown in the table some SNVs are very close to each other means they have a high probability of being carried together. Especially rs4303794, rs11088551, and rs61299115 have a score of 1. The linkage shows that they are in perfect linkage and are expected to be carried almost always together.

In addition to the R^2 matrix shown above, a heatmap has been created, the illustrated version of the same map. The following heatmap shows the linkages in colors. The darker the color, the stronger the linkage.

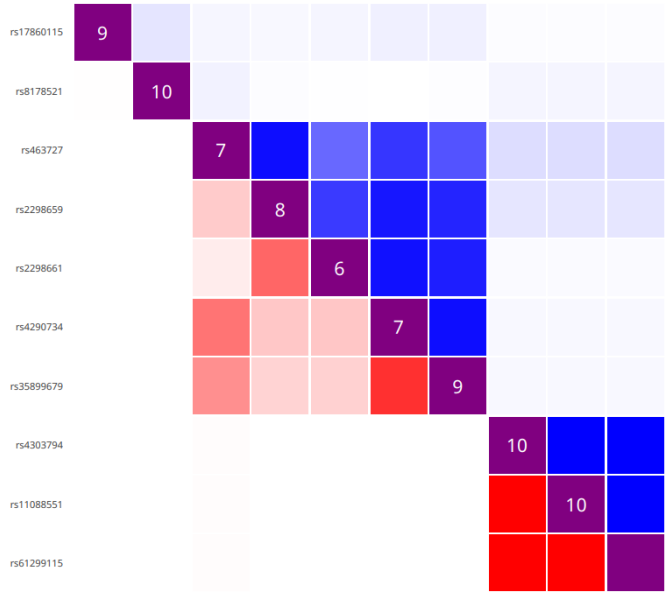

Figure 3 Linkage matrix heatmap for chr 21 variants.

4.7.KEGG Pathway Analysis

With SNVs that resulted in significant differences in frequency between non-finnish European population and Turkish population, a KEGG pathway analysis has been conducted. Figure 3 and Figure 4 show results of the KEGG pathway analysis.

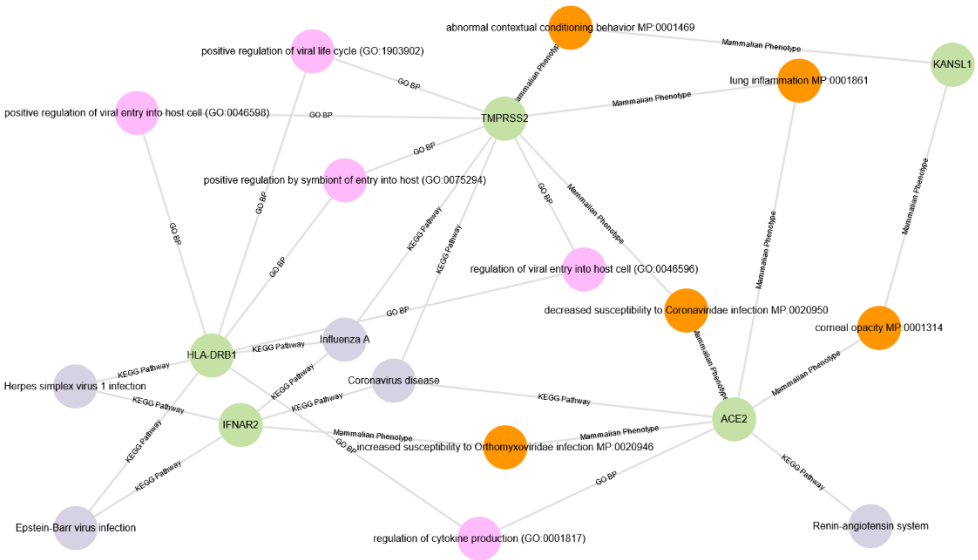

Figure 4 KEGG pathway (enrichr)

The subnetwork reveals significant associations related to COVID-19 and other viral diseases. Gene Ontology data shows that HLA-DRB1 and TMPRSS2 are involved in the biological processes regulating viral entry into host cells and positively regulating viral entry into host cells. Both genes also play roles in the positive regulation of viral life cycles and the positive regulation by symbionts of entry into host cells. Additionally, HLA-DRB1 and ACE2 are implicated in the regulation of cytokine production, a critical response during viral infections.

From the MGI Mammalian Phenotype database, knockout (KO) mice for IFNAR2 and ACE2 showed increased susceptibility to Orthomyxoviridae infections. TMPRSS2 and ACE2 KO mice exhibited lung inflammation, and KANSL1 and ACE2 KO mice displayed corneal opacity. Interestingly, TMPRSS2 and ACE2 KO mice demonstrated decreased susceptibility to Coronaviridae infections, while TMPRSS2 and KANSL1 KO mice showed abnormal contextual conditioning behavior.

From the KEGG pathway database, the gene products HLA-DRB1 and IFNAR2 are part of the Epstein-Barr virus infection pathway. HLA-DRB1, IFNAR2, and TMPRSS2 are linked to the Influenza A pathway. HLA-DRB1 and IFNAR2 are associated with the Herpes simplex virus 1 infection pathway, while IFNAR2, TMPRSS2, and ACE2 are part of the Coronavirus disease pathway. Additionally, ACE2 is involved in the Renin-angiotensin system pathway, which has implications for COVID-19 severity and progression.

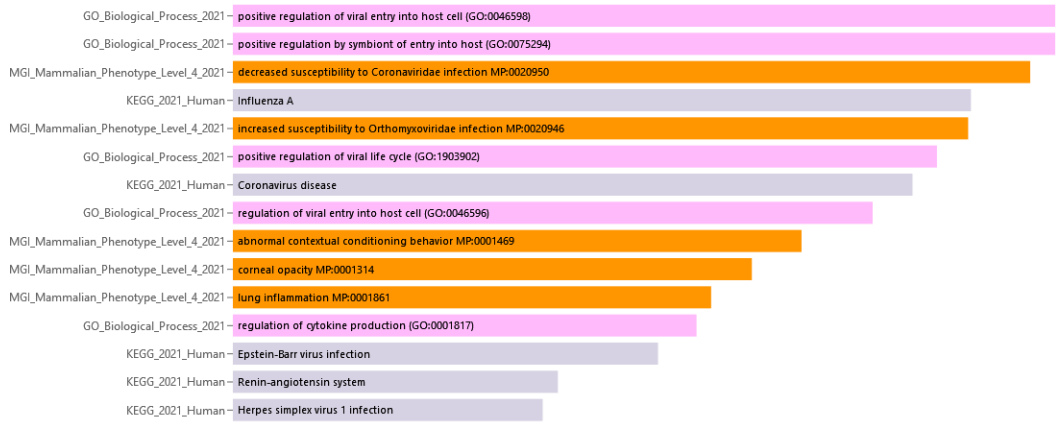

Figure 5 KEGG Pathway bar chart model (enrichr)

CHAPTER 5

5. DISCUSSION

This study presents a detailed genetic analysis of Single Nucleotide Variants (SNVs) and their association with COVID-19 susceptibility among patients at Hacettepe Hospital in Türkiye. Through a robust methodological approach encompassing sample collection, DNA isolation, PCR, sequencing, and advanced statistical analyses, our findings reveal significant insights into the genetic predispositions that may influence COVID-19 outcomes within the Turkish population. This discussion seeks to contextualize these results within the broader genetic research landscape and public health implications related to COVID-19.

The genetic makeup of the contemporary Turkish population exhibits notable similarities with non-Finnish European populations, particularly those in the Mediterranean region, as evidenced by clustering with Iberians from Spain and Tuscans from Italy. This clustering suggests a shared genetic heritage, further supported by the similar frequency distributions of GWAS (Genome-Wide Association Study) SNPs between Turkish and European populations, which indicate a greater proportion of ancestry sharing. Additionally, SNPs associated with cholesterol levels display higher frequencies in the Turkish population, aligning with the observed lower total cholesterol counts and different lipid profiles compared to Western Europeans. This study assumes that the population of Turkey shares a genetic predisposition similar to that of the non-Finnish European population under normal conditions. In other words, it is thought that the average person in Turkey and the average MAF in Europe populations show almost the same genetic diversity. The European population mentioned here, and the European population mentioned hereinafter, is kept separate from the Finnish population. From now on, what will be referred to as the European population in this discussion section is actually the non-Finnish European population (Alkan et al., 2014; Özçelik et al., 2010).

During the frequency comparisons in the study, significant allele frequency differences were observed for rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917 variants when the study population and the EU population were compared. The rs17860115 variant appeared to show higher frequency in the study population than in Europe. It has been observed that the rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917 variants are seen with higher frequency in the European population. It is possible to consider potential inferences, considering

that the entire study cohort was admitted to the hospital due to the disease, which indicates COVID-19 susceptibility.

It is conceivable that one reason why the rs17860115 variant has appeared more frequently in the patient population is that this mutation increases the risk of contracting COVID-19 disease. For other variants (rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917), it can be thought that these variants reduce the likelihood of contracting COVID-19 as they appear more frequently in the general population.

In the bootstrap analysis variants rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917 showed differences than the general population. Our data was simulated with 100 alleles for 1000 cohorts in the bootstrap algorithm. After 1000 cohorts were simulated with 100 alleles with the Bootstrap algorithm, 95 percent confidence interval limits were determined. If the allele frequency seen in the European population was not within this confidence interval, it was defined as having a significant difference. It is reasonable to think that the frequencies of these variants, which differ in 1000 simulated cohorts containing 100 alleles, significantly differ between Turkey and Europe. The variants that showed differences in the z-test and those that showed differences in the bootstrap analysis completely overlapped, yielding the same direction. Variants are different, and the way they differ, whether they are high or low, are in the same direction.

It was observed that the genes with the variants that showed differences were ACE2, FBRSL1, HLA-DRB1, IFNAR2, KANSL1, and TMPRSS2 genes. In studies examining the relationship of the ACE2 gene with COVID-19, direct or inversely proportional relationships were observed between ACE2 gene polymorphisms and COVID-19 severity (Rodrigues & de Oliveira, 2021; Karahalil & Elkama, 2020). Since the scope of this study is not the severity of COVID-19 but its susceptibility, it can be thought that the rs61299115 variant, which is on the ACE2 gene seen in this study, potentially reduces the risk of contracting COVID-19.

The FBRSL1 gene has not been directly studied concerning COVID-19. However, the rs56106917 variant inspected in this study shows a potential relation with the disease.

Some studies have shown potential relationships between variants in the HLA-DRB1 gene and COVID-19 disease severity (Anzurez et al. 2021). In this study, the rs9271609 variant on the HLA-DRB1 gene may be potentially associated with decreased disease risk.

Research has consistently identified the IFNAR2 gene as a significant factor in COVID-19 susceptibility and severity (Ma et al. 2020; Ma et al. 2021). With conformity, in this research, the rs17860115 variant is potentially related to COVID-19 susceptibility. According to the data from dbSNP, the single nucleotide

polymorphism rs17860115 is located within the 5' untranslated region (UTR) of the gene. The positioning of this variant in the UTR suggests that it may play a significant role in the regulation of gene expression or translation (Brown, 2002).

The KANSL1 gene, involved in chromatin remodeling and gene expression, has not been directly linked to COVID-19 in the literature. However, there is evidence that genetic factors, including the NKG2C receptor encoding the KLRC2 gene and HLA-E variants, can influence the severity of COVID-19 (Vietzen et al., 2021). In this, the variant rs2532300 on the KANSL1 gene has been found to be potentially related to decreased COVID-19 susceptibility.

The TMPRSS2 gene has been found to play a significant role in COVID-19 susceptibility and severity. Rokni et al. (2022) identified an increased risk of COVID-19 in carriers of certain TMPRSS2 polymorphisms, while Wulandari et al. (2021) found a possible association between the p.Val160Met polymorphism and SARS-CoV-2 infectivity and disease outcome. rs2298659, rs2298661, rs2532300, rs34624090, and rs4290734 variants in this study showed a potentially protective effect towards COVID-19.

In this study, we did not observe any of the following variants: rs2532300, rs34624090, rs61299115, and rs56106917. In other words, these variants were not encountered even once in the cohort of 120 patients, which means 240 alleles. It is a matter of curiosity that these variants, known to be encountered with frequencies between 20 percent and 49 percent in the European population, are not observed in the COVID-19 cohort. Any discussion on these variants requires more detailed studies conducted with larger sample groups in the future.

It was mentioned that the rs4303794, rs11088551, and rs61299115 variants show perfect linkage features. This means that if a person has a polymorphism at one of these variant locations, they also have a polymorphism at the other location. The MAF for the European population for alleles in these locations are 0.422, 0.422, and 0.421, as expected for variants with perfect linkage. In other words, there is a difference of one thousandth. Also, when we look at the Turkish population, the rs4303794 and rs11088551 variants show a frequency of 0.404 (they are close to each other with the European frequency, and there is no significant difference), while the frequency of the rs61299115 variant is measured as 0. More comprehensive studies with a larger sample size are required to explain the reason for this situation entirely.

Identifying unique SNVs and their linkage patterns offers new avenues for research and potential targets for therapeutic intervention. As the global community continues to combat COVID-19, integrating genetic research into public health strategies remains a priority, promising more effective responses to this and future pandemics. Further research in this area will enhance our understanding of the genetic basis of infectious diseases and improve our ability to predict and mitigate their impacts on diverse populations.

Despite the strengths of our study, there are several limitations to consider. While adequate for initial analyses, the sample size limits our findings' generalizability across the broader Turkish population. Future studies with larger, more diverse cohorts must validate these results and refine the identified genetic markers.

Additionally, while our study focused on genetic predispositions, the interaction between genetic, environmental, and social factors is crucial in determining disease outcomes. Comprehensive models integrating these factors are needed to provide a more complete picture of COVID-19 susceptibility and severity.

CHAPTER 6

6. CONCLUSION

In conclusion, our comprehensive analysis of SNVs and their association with COVID-19 susceptibility and severity among the Hacettepe Hospital cohort, assumed to represent the Turkish population, has provided valuable insights. Due to the small sample size, we have proposed a bootstrapping-based analysis method to observe the standard deviation within our samples, which is utilized during the maf percentage comparison between our cohort and the European population.

The study highlights significant allele frequency variations between the Turkish cohort and the non-Finnish European population, particularly in SNVs such as rs17860115, rs2298659, rs2298661, rs4290734, rs9271609, rs2532300, rs34624090, rs61299115, and rs56106917. These differences suggest potential genetic predispositions that could influence COVID-19 outcomes. Moreover, the linkage analysis revealed strong correlations between specific SNVs, indicating that these genetic loci may be inherited together, which could affect disease susceptibility and severity.

The findings underscore the importance of considering genetic variability within and between populations in public health strategies, particularly in the context of infectious diseases like COVID-19. They also highlight the need for further research with larger, more diverse cohorts to confirm these associations and to explore the potential mechanisms underlying these genetic influences. By expanding our understanding of genetic factors in disease susceptibility and severity, we can better tailor interventions and improve outcomes for affected populations. Thus, while promising, the results presented should be viewed as a stepping stone towards more extensive genetic research that could ultimately inform more effective public health responses to COVID-19 and other infectious diseases.

REFERENCES

Alkan, C., Kavak, P., Somel, M., Gokcumen, O., Ugurlu, S., Saygi, C., Dal, E., Bugra, K., Güngör, T., Sahinalp, S. C., Özören, N., & Bekpen, C. (2014). Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC genomics*, *15*(1), 963. https://doi.org/10.1186/1471-2164-15-963

Alkharsah, K. R., Aljaroodi, S. A., Rahman, J. U., Alnafie, A. N., Al Dossary, R., Aljindan, R. Y., Alnimr, A. M., & Hussen, J. (2022). Low levels of soluble DPP4 among Saudis may have constituted a risk factor for MERS endemicity. PloS One, 17(4), e0266603. https://doi.org/10.1371/journal.pone.0266603

Andolfo, I., Russo, R., Lasorsa, V. A., Cantalupo, S., Rosato, B. E., Bonfiglio, F., Frisso, G., Abete, P., Cassese, G. M., Servillo, G., Esposito, G., Gentile, I., Piscopo, C., Villani, R., Fiorentino, G., Cerino, P., Buonerba, C., Pierri, B., Zollo, M., … Capasso, M. (2021). Common variants at 21q22.3 locus influence MX1 and TMPRSS2 gene expression and susceptibility to severe COVID-19. IScience, 24(4). https://doi.org/10.1016/j.isci.2021.102322

Angulo-Aguado, M., Carrillo-Martinez, J. C., Contreras-Bravo, N. C., Morel, A., Parra-Abaunza, K., Usaquén, W., Fonseca-Mendoza, D. J., & Ortega-Recalde, O. (2024). Next-generation sequencing of host genetics risk factors associated with COVID-19 severity and long-COVID in Colombian population. Scientific Reports, 14(1), 8497. https://doi.org/10.1038/s41598-024-57982-3

Anzurez, A., Naka, I., Miki, S., Nakayama-Hosoya, K., Isshiki, M., Watanabe, Y., Nakamura-Hoshi, M., Seki, S., Matsumura, T., Takano, T., Onodera, T., Adachi, Y., Moriyama, S., Terahara, K., Tachikawa, N., Yoshimura, Y., Sasaki, H., Horiuchi, H., Miyata, N., … Kawana-Tachikawa, A. (2021). Association of HLA-DRB1*09:01 with severe COVID-19. HLA, 98(1), 37–42. https://doi.org/https://doi.org/10.1111/tan.14256

ARAL, N., & BAKIR, H. (2022). Spatiotemporal Analysis of Covid-19 in Turkey. Sustainable Cities and Society, 76, 103421. https://doi.org/https://doi.org/10.1016/j.scs.2021.103421

Arredondo-Alonso, S., Pöntinen, A. K., Cléon, F., Gladstone, R. A., Schürch, A. C., Johnsen, P. J., Samuelsen, Ø., & Corander, J. (2021). A high-throughput

multiplexing and selection strategy to complete bacterial genomes. GigaScience, 10(12). https://doi.org/10.1093/gigascience/giab079

Asselta, R., Paraboschi, E. M., Mantovani, A., & Duga, S. (2020). ACE2 and TMPRSS2 variants and expression as candidates to sex and country differences in COVID-19 severity in Italy. Aging, 12(11), 10087–10098. https://doi.org/10.18632/aging.103415

Azzarà, A., Cassano, I., Paccagnella, E., Tirindelli, M. C., Nobile, C., Schittone, V., Lintas, C., Sacco, R., & Gurrieri, F. (2022). Genetic variants determine intrafamilial variability of SARS-CoV-2 clinical outcomes in 19 Italian families. PloS One, 17(10), e0275988. https://doi.org/10.1371/journal.pone.0275988

Barash, A., Machluf, Y., Ariel, I., & Dekel, Y. (2020). The Pursuit of COVID-19 Biomarkers: Putting the Spotlight on ACE2 and TMPRSS2 Regulatory Sequences. Frontiers in Medicine, 7. https://doi.org/10.3389/fmed.2020.582793

Beck, S. (1993). Multiplex DNA sequencing. Methods in Molecular Biology (Clifton, N.J.), 23, 225–234. https://doi.org/10.1385/0-89603-248-5:225

Bostan, S., Erdem, R., Öztürk, Y. E., Kılıç, T., & Yılmaz, A. (2020). The Effect of COVID-19 Pandemic on the Turkish Society. *Electronic Journal of General Medicine, 17*(6), em237. https://doi.org/10.29333/ejgm/7944

Brody, T. (2016). Chapter 19 - Biomarkers. In T. Brody (Ed.), Clinical Trials (Second Edition) (Second Edition, pp. 377–419). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-804217-5.00019-9

Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002.https://www.ncbi.nlm.nih.gov/books/NBK21136/

Brown, T. (n.d.). Brown T 1957.PDF.

Callaway, E. (2021). Genetic variants linked to COVID risk. Nature, 595, 346–348. https://doi.org/10.1038/s41586-021-03767-x

Cappadona, C., Rimoldi, V., Paraboschi, E. M., & Asselta, R. (2023). Genetic susceptibility to severe COVID-19. Infection, Genetics and Evolution, 110, 105426. https://doi.org/https://doi.org/10.1016/j.meegid.2023.105426

Centers for Disease Control and Prevention. (2024, April 9). *About COVID-19*. https://www.cdc.gov/coronavirus/2019-ncov/your-health/about-covid-19.html

Choudhary, S., Sreenivasulu, K., Mitra, P., Misra, S., & Sharma, P. (2021). Role of Genetic Variants and Gene Expression in the Susceptibility and Severity of COVID-19. Alm, 41(2), 129–138. https://doi.org/10.3343/alm.2021.41.2.129

Chuang, L.-Y., Cheng, Y.-H., & Yang, C.-H. (2013). Specific primer design for the polymerase chain reaction. Biotechnology Letters, 35(10), 1541–1549. https://doi.org/10.1007/s10529-013-1249-8

Church, G. M., & Kieffer-Higgins, S. (1988). Multiplex DNA sequencing. Science (New York, N.Y.), 240(4849), 185–188. https://doi.org/10.1126/science.3353714

Clark, D. P., Pazdernik, N. J., & McGehee, M. R. (2019). Chapter 6 - Polymerase Chain Reaction. In D. P. Clark, N. J. Pazdernik, & M. R. McGehee (Eds.), Molecular Biology (Third Edition) (Third Edition, pp. 168–198). Academic Cell. https://doi.org/https://doi.org/10.1016/B978-0-12-813288-3.00006-9

Dasgupta, I., Saini, S., Khan, M. A., & Chaudhary, K. (2023). 1 - Population-level differences in COVID-19 prevalence, severity, and clinical outcome. In R. Pandey (Ed.), Genomic Surveillance and Pandemic Preparedness (pp. 3–25). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-443-18769-8.00008-8

Dieffenbach, C. W., Lowe, T. M., & Dveksler, G. S. (1993). General concepts for PCR primer design. PCR Methods and Applications, 3(3), S30-7. https://doi.org/10.1101/gr.3.3.s30

Drabik, A., Bodzoń-Kułakowska, A., & Silberring, J. (2016). 7 - Gel Electrophoresis. In P. Ciborowski & J. Silberring (Eds.), Proteomic Profiling and Analytical Chemistry (Second Edition) (Second Edition, pp. 115–143). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-63688-1.00007-0

El-Maradny, Y. A., Ramadan, A. A., Chavda, V. P., Balar, P. C., & Redwan, E. M. (2024). Chapter 22 - The fast-track development of COVID-19 vaccines. In V. P. Chavda, L. K. Vora, & V. Apostolopoulos (Eds.), Advanced Vaccination Technologies for Infectious and Chronic Diseases (pp. 415–440). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-443-18564-9.00027-8

Garbieri, T. F., Brozoski, D. T., Dionísio, T. J., Santos, C. F., & Neves, L. T. das. (2017). Human DNA extraction from whole saliva that was fresh or stored for 3, 6 or 12 months using five different protocols. Journal of Applied Oral Science : Revista FOB, 25(2), 147–158. https://doi.org/10.1590/1678-77572016-0046

Gudiseva, H. V, Hansen, M., Gutierrez, L., Collins, D. W., He, J., Verkuil, L. D., Danford, I. D., Sagaser, A., Bowman, A. S., Salowe, R., Sankar, P. S., Miller-Ellis, E., Lehman, A., & O'Brien, J. M. (2016). Saliva DNA quality and genotyping efficiency in a predominantly elderly population. BMC Medical Genomics, 9, 17. https://doi.org/10.1186/s12920-016-0172-y

Gunderson, K., Barker, D. L., Bibikova, M., & Fan, J.-B. (2007). CHAPTER 7 - Genotype and Epigenotype by Single Nucleotide Variant (SNV) Analysis. In J. F. Loring, R. L. Wesselschmidt, & P. H. Schwartz (Eds.), Human Stem Cell Manual (pp. 85–95). Academic Press. https://doi.org/https://doi.org/10.1016/B978-012370465-8/50012-0

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hossain, M. S., Tonmoy, M. I. Q., Fariha, A., Islam, M. S., Roy, A. S., Islam, M. N., Kar, K., Alam, M. R., & Rahaman, M. M. (2021). Prediction of the Effects of Variants and Differential Expression of Key Host Genes ACE2, TMPRSS2, and FURIN in SARS-CoV-2 Pathogenesis: An In Silico Approach. Bioinformatics and Biology Insights, 15, 11779322211054684. https://doi.org/10.1177/11779322211054684

Hughes, G. M., Gang, L., Murphy, W. J., Higgins, D. G., & Teeling, E. C. (2013). Using Illumina next generation sequencing technologies to sequence multigene families in de novo species. Molecular Ecology Resources, 13(3), 510–521. https://doi.org/10.1111/1755-0998.12087

Iyer, G. R., Samajder, S., Zubeda, S., S, D. S. N., Mali, V., Pv, S. K., Sharma, A., Abbas, N. Z., Bora, N. S., Narravula, A., & Hasan, Q. (2020). Infectivity and Progression of COVID-19 Based on Selected Host Candidate Gene Variants. Frontiers in Genetics, 11, 861. https://doi.org/10.3389/fgene.2020.00861

Karahalil, B., & Elkama, A. (2020). The Impact of ACE2 Gene Polymorphism in the Development of COVID-19 Disease. Gazi Medical Journal, 31, 518–522. https://doi.org/http://dx.doi.org/10.12996/gmj.2020.122

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. Genome Medicine, 12(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Kousathanas, A., Pairo-Castineira, E., Rawlik, K., Stuckey, A., Odhams, C. A., Walker, S., Russell, C. D., Malinauskas, T., Wu, Y., Millar, J., Shen, X., Elliott, K. S., Griffiths, F., Oosthuyzen, W., Morrice, K., Keating, S., Wang, B., Rhodes, D., Klaric, L., … team, R. I. of E. (2022). Whole-genome sequencing reveals host factors underlying critical COVID-19. Nature, 607(7917), 97–103. https://doi.org/10.1038/s41586-022-04576-6

Leal, V. N. C., Paulino, L. M., Cambui, R. A. G., Zupelli, T. G., Yamada, S. M., Oliveira, L. A. T., Dutra, V. de F., Bub, C. B., Sakashita, A. M., Yokoyama, A. P. H., Kutner, J. M., Vieira, C. A., Santiago, W. M. de S., Andrade, M. M. S., Teixeira, F. M. E., Alberca, R. W., Gozzi-Silva, S. C., Yendo, T. M., Netto, L. C., … Pontillo, A. (2023). A common variant close to the “tripwire” linker region of NLRP1 contributes to severe COVID-19. Inflammation Research, 72(10), 1933–1940. https://doi.org/10.1007/s00011-022-01670-3

Lee, P. Y., Costumbrado, J., Hsu, C.-Y., & Kim, Y. H. (2012). Agarose gel electrophoresis for the separation of DNA fragments. Journal of Visualized Experiments : JoVE, 62. https://doi.org/10.3791/3923

López-Bielma, M. F., Falfán-Valencia, R., Abarca-Rojano, E., & Pérez-Rubio, G. (2023). Participation of Single-Nucleotide Variants in IFNAR1 and IFNAR2 in the Immune Response against SARS-CoV-2 Infection: A Systematic Review. In Pathogens (Vol. 12, Issue 11). https://doi.org/10.3390/pathogens12111320

Louten, J. (2023). Chapter 7 - Virology research and diagnosis of viral infections. In J. Louten (Ed.), Essential Human Virology (Second Edition) (Second Edition, pp. 119–144). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-90565-7.00007-1

Ma, Y., Huang, Y., Zhao, S., Yao, Y., Zhang, Y., Qu, J., Wu, N., & Su, J. (2020). Integrative Genomics Analysis Reveals a Novel 21q22.11 Locus Contributing to Susceptibility of COVID-19. MedRxiv. https://api.semanticscholar.org/CorpusID:221778689

Ma, Y., Huang, Y., Zhao, S., Yao, Y., Zhang, Y., Qu, J., Wu, N., & Su, J. (2021). Integrative genomics analysis reveals a 21q22.11 locus contributing risk to COVID-19. Human Molecular Genetics, 30(13), 1247–1258. https://doi.org/10.1093/hmg/ddab125

Mwendwa, F., Kanji, A., Bukhari, A. R., Khan, U., Sadiqa, A., Mushtaq, Z., Nasir, N., Mahmood, S. F., Aamir, U. B., & Hasan, Z. (2024). Shift in SARS-CoV-2 variants of concern from Delta to Omicron was associated with reduced hospitalizations, increased risk of breakthrough infections but lesser disease severity. Journal of Infection and Public Health, 17(6), 1100–1107. https://doi.org/https://doi.org/10.1016/j.jiph.2024.04.025

National Center for Biotechnology Information. (2024a, May 14). ACE2 angiotensin converting enzyme 2 [Homo sapiens (human)]. Gene ID: 59272. Retrieved from https://www.ncbi.nlm.nih.gov/gene/59272

National Center for Biotechnology Information. (2024b, May 5). BCL11A BCL11 transcription factor A [Homo sapiens (human)]. Gene ID: 53335. Retrieved from https://www.ncbi.nlm.nih.gov/gene/53335

National Center for Biotechnology Information. (2024c, April 30). DPP4 dipeptidyl peptidase 4 [Homo sapiens (human)]. Gene ID: 1803. Retrieved from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=1803

National Center for Biotechnology Information. (2024d, May 13). DPP9 dipeptidyl peptidase 9 [Homo sapiens (human)]. Gene ID: 91039. Retrieved from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=91039

National Center for Biotechnology Information. (2024e, April 3). ELF5 E74 like ETS transcription factor 5 [Homo sapiens (human)]. Gene ID: 2001. Retrieved from https://www.ncbi.nlm.nih.gov/gene/2001

National Center for Biotechnology Information. (2024f, April 3). FBRSL1 fibrosin like 1 [Homo sapiens (human)]. Gene ID: 57666. Retrieved from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=57666

National Center for Biotechnology Information. (2024g, April 7). HLA-DRB1 major histocompatibility complex, class II, DR beta 1 [Homo sapiens (human)]. Gene ID: 3123. Retrieved from https://www.ncbi.nlm.nih.gov/gene/3123

National Center for Biotechnology Information. (2024h, April 11). IFNAR2 interferon alpha and beta receptor subunit 2 [Homo sapiens (human)]. Gene ID: 3455. Retrieved from https://www.ncbi.nlm.nih.gov/gene/3455

National Center for Biotechnology Information. (2024i, May 2). IL10RB interleukin 10 receptor subunit beta [Homo sapiens (human)]. Gene ID: 3588. Retrieved from https://www.ncbi.nlm.nih.gov/gene/3588

National Center for Biotechnology Information. (2024j, May 3). KANSL1 KAT8 regulatory NSL complex subunit 1 [Homo sapiens (human)]. Gene ID: 284058. Retrieved from https://www.ncbi.nlm.nih.gov/gene/284058

National Center for Biotechnology Information. (2024k, May 6). TMPRSS2 transmembrane serine protease 2 [Homo sapiens (human)]. Gene ID: 7113. Retrieved from https://www.ncbi.nlm.nih.gov/gene/7113

Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R., & Braun, A. (2004). Large-scale validation of Single Nucleotide Variants in gene regions. Genome Research, 14(8), 1664–1668. https://doi.org/10.1101/gr.2421604

Niemi, M. E. K., Karjalainen, J., Liao, R. G., Neale, B. M., Daly, M., Ganna, A., Pathak, G. A., Andrews, S. J., Kanai, M., Veerapen, K., Fernandez-Cadenas, I., Schulte, E. C., Striano, P., Marttila, M., Minica, C., Marouli, E., Karim, M. A., Wendt, F. R., Savage, J., … leader, A. team. (2021). Mapping the human genetic architecture of COVID-19. Nature, 600(7889), 472–477. https://doi.org/10.1038/s41586-021-03767-x

Özen, F. (2024). Random forest regression for prediction of Covid-19 daily cases and deaths in Turkey. Heliyon, 10(4), e25746. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e25746

Özçelik, T., Kanaan, M., Avraham, K. B., Yannoukakos, D., Mégarbané, A., Tadmouri, G. O., Middleton, L., Romeo, G., King, M. C., & Levy-Lahad, E. (2010). Collaborative genomics for human health and cooperation in the

Mediterranean region. *Nature genetics*, *42*(8), 641–645. https://doi.org/10.1038/ng0810-641

Phalke, S., Sawant, S. A., Samudra, P., Yadav, P., Chakraborty, C., Jadhav, A., & Nandi, S. S. (2024). Chapter 5.1 - Viral Genome Sequencing and Its Significance in Latest Clinical and Research Findings. In S. Das & H. R. Dash (Eds.), Microbial Diversity in the Genomic Era (Second Edition) (Second Edition, pp. 517–539). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-443-13320-6.00001-9

Posadas-Sánchez, R., Fragoso, J. M., Sánchez-Muñoz, F., Rojas-Velasco, G., Ramírez-Bello, J., López-Reyes, A., Martínez-Gómez, L. E., Sierra-Fernández, C., Rodríguez-Reyna, T., Regino-Zamarripa, N. E., Ramírez-Martínez, G., Zuñiga-Ramos, J., & Vargas-Alarcón, G. (2022). Association of the Transmembrane Serine Protease-2 (TMPRSS2) Polymorphisms with COVID-19. Viruses, 14(9). https://doi.org/10.3390/v14091976

Posadas-Sánchez, R., Sánchez-Muñoz, F., Guzmán-Martín, C. A., Hernández-Díaz Couder, A., Rojas-Velasco, G., Fragoso, J. M., & Vargas-Alarcón, G. (2021). Dipeptidylpeptidase-4 levels and DPP4 gene polymorphisms in patients with COVID-19. Association with disease and with severity. Life Sciences, 276, 119410. https://doi.org/10.1016/j.lfs.2021.119410

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., & Lander, E. S. (2001). Linkage disequilibrium in the human genome. Nature, 411(6834), 199–204. https://doi.org/10.1038/35075590

Rodrigues, R., & Costa de Oliveira, S. (2021). The Impact of Angiotensin-Converting Enzyme 2 (ACE2) Expression Levels in Patients with Comorbidities on COVID-19 Severity: A Comprehensive Review. In Microorganisms (Vol. 9, Issue 8). https://doi.org/10.3390/microorganisms9081692

Rogers, N. L., Cole, S. A., Lan, H.-C., Crossa, A., & Demerath, E. W. (2007). New saliva DNA collection method compared to buccal cell collection techniques for epidemiological studies. American Journal of Human Biology : The Official Journal of the Human Biology Council, 19(3), 319–326. https://doi.org/10.1002/ajhb.20586

Rokni, M., Heidari Nia, M., Sarhadi, M., Mirinejad, S., Sargazi, S., Moudi, M., Saravani, R., Rahdar, S., & Kargar, M. (2022). Association of TMPRSS2 Gene Polymorphisms with COVID-19 Severity and Mortality: a Case-Control Study with Computational Analyses. Applied Biochemistry and Biotechnology, 194(8), 3507–3526. https://doi.org/10.1007/s12010-022-03885-w

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNV: the NCBI database of genetic variation. Nucleic Acids Research, 29(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Spiteri, G., Fielding, J., Diercke, M., Campese, C., Enouf, V., Gaymard, A., Bella, A., Sognamiglio, P., Sierra Moros, M. J., Riutort, A. N., Demina, Y. V, Mahieu, R., Broas, M., Bengnér, M., Buda, S., Schilling, J., Filleul, L., Lepoutre, A., Saura, C., … Ciancio, B. C. (2020). First cases of coronavirus disease 2019 (COVID-19) in the WHO European Region, 24 January to 21 February 2020. Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin, 25(9). https://doi.org/10.2807/1560-7917.ES.2020.25.9.2000178

Vietzen, H., Zoufaly, A., Traugott, M., Aberle, J., Aberle, S. W., & Puchhammer-Stöckl, E. (2021). Deletion of the NKG2C receptor encoding KLRC2 gene and HLA-E variants are risk factors for severe COVID-19. Genetics in Medicine, 23(5), 963–967. https://doi.org/10.1038/s41436-020-01077-7

Viruses, C. S. G. of the I. C. on T. of. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology, 5(4), 536–544. https://doi.org/10.1038/s41564-020-0695-z

Wages, J. M. (2005). POLYMERASE CHAIN REACTION. In P. Worsfold, A. Townshend, & C. Poole (Eds.), Encyclopedia of Analytical Science (Second Edition) (Second Edition, pp. 243–250). Elsevier. https://doi.org/https://doi.org/10.1016/B0-12-369397-7/00475-1

Wooster, L., Nicholson, C. J., Sigurslid, H. H., Cardenas, C. L. L., & Malhotra, R. (2020). Polymorphisms in the ACE2 Locus Associate with Severity of COVID-19 Infection. MedRxiv. https://doi.org/10.1101/2020.06.18.20135152

Wulandari, L., Hamidah, B., Pakpahan, C., Damayanti, N. S., Kurniati, N. D., Adiatmaja, C. O., Wigianita, M. R., Soedarsono, Husada, D., Tinduh, D., Prakoeswa, C. R. S., Endaryanto, A., Puspaningsih, N. N. T., Mori, Y., Lusida, M. I., Shimizu, K., & Oceandy, D. (2021). Initial study on TMPRSS2 p.Val160Met genetic variant in COVID-19 patients. Human Genomics, 15(1), 29. https://doi.org/10.1186/s40246-021-00330-7

Zhao, X., Wu, X., Xiao, J., Zhang, L., Hao, Y., Xiao, C., Zhang, B., Li, J., & Jiang, X. (2022). Are female-specific cancers long-term sequelae of COVID-19? Evidence from a large-scale genome-wide cross-trait analysis. MedRxiv. https://doi.org/10.1101/2022.08.25.22279195

Zverinova, S., & Guryev, V. (2022). Variant calling: Considerations, practices, and developments. Human Mutation, 43(8), 976–985. https://doi.org/10.1002/humu.24311

APPENDICES

A. Nanodrop Results

Table 5 Nanodrop results to investigate DNA products in isolates

Sample No	Nucleic Acid(ng/uL)	A260/A280	A260/A230	A260	A280
1	6,174	1.497	-442	123	82
2	69,901	1.639	1.622	1.398	853
3	67,300	1.668	1.579	1.346	807
4	58,782	1.689	2.319	1.176	696
5	82,480	1.495	886	1.65	1.103
6	61,726	1.709	775	1.235	722
7	6,161	1.258	-3.207	123	98
10	3,550	1.258	-185	71	56
11	13,619	1.516	-9.272	272	0.18
12	112,768	1.641	1.222	2.255	1.374
13	145,046	1.774	2.15	2.901	1.635
14	116,280	1.729	2.039	2.326	1.345
16	5,531	1.282	-0.49	111	86
17	110,348	1.707	1.743	2.207	1.293
19	133,212	1.601	1.091	2.664	1.664
20	34,434	1.576	1.627	689	437
21	73,139	1.702	2.147	1.463	0.86
22	291,421	1.772	1.724	5.828	3.289
23	70,966	1.669	1.753	1.419	0.85
24	32,314	1.614	1.676	646	0.4
25	49,955	1.716	3.049	999	582
26	35,558	1.605	1.206	711	443
27	14,033	1.784	-2.152	281	157
28	38,017	1.673	2.464	0.76	455
29	10,451	1.548	-1.303	209	135
31	38,959	1.708	5.181	779	456
32	25,989	1.498	1.681	0.52	347
33	107,004	1.771	2.379	2.14	1.209
35	15,349	1.63	-10.239	307	188

Sample No	Nucleic Acid(ng/uL)	A260/A280	A260/A230	A260	A280
36	11,215	1.489	-1.54	224	151
38	49,946	1.64	1.853	999	609
39	17,434	1.703	-4.992	349	205
40	25,999	1.586	3.367	0.52	328
41	214,403	1.737	1.464	4.288	2.468
42	42,200	1.583	2.056	844	533
43	132,360	1.759	2.067	2.647	1.505
44	17,590	1.567	12.617	352	225
45	6,612	1.375	-837	132	96
46	25,176	1.714	-6.944	504	294
47	174,012	1.798	2.827	3.48	1.936
48	3,790	1.224	-358	76	62
49	5,078	1.059	-381	102	96
50	44,194	1.77	-80.181	884	499
51	294,288	1.546	866	5.886	3.806
52	24,002	1.661	-1.100.688	0.48	289
54	10,252	1.551	-687	205	132
55	43,885	1.676	2.662	878	524
56	44,789	1.749	266.211	896	512
58	137,535	1.709	1.541	2.751	1.609
59	48,839	1.604	3.364	977	609
60	125,820	1.73	2.115	2.516	1.454
61	106,076	1.791	2.611	2.122	1.185
62	5,192	1.262	-396	104	82
66	26,451	1.577	3.215	529	335
67	44,394	1.761	-18.231	341	194
68	190,020	1.701	1.38	3.8	2.234
69	123,188	1.777	2.635	2.464	1.386
70	9,075	1.515	-943	182	0.12
71	18,662	1.454	3.257	373	257
73	6,821	1.279	-1.021	136	107
75	12,728	1.394	-2.383	255	183
76	40,087	1.634	2.007	802	491
77	24,027	1.668	10.234	481	288
78	3,001	1.116	-206	0.06	54
79	18,920	1.663	-8.49	378	228
81	69,180	1.687	1.841	1.384	0.82
82	22,124	1.632	4.993	442	271
83	18,476	1.457	6.693	0.37	254

Sample No	Nucleic Acid(ng/uL)	A260/A280	A260/A230	A260	A280
84	5,145	1.333	-385	103	77
85	49,051	593	41	981	1.654
86	6,105	44.501,00	-584	122	0.11
88	21,992	1.705	-22.032	0.44	258
89	164,600	1.517	789	3.292	2.17
90	92,575	1.729	1.884	1.852	1.071
91	37,980	1.607	1.85	0.76	473
92	174,952	1.625	01.01	3.499	2.153
93	106,832	1.752	2.641	2.137	1.219
93	61,965	1.699	2.935	1.239	0.73
94	299,177	1.843	2.79	5.984	3.247
96	10,628	1.415	567	213	0.15
97	7,485	1.374	-291	0.15	109
102	175,429	1.674	1.158	3.509	2.096
103	68,440	1.713	1.84	1.369	799
105	5,195	1.362	-482	104	76
106	41,416	1.602	1.485	828	517
107	6,432	1.333	-759	129	97
108	8,417	1.472	-922	168	114
109	24,667	1.656	-19.796	493	298
111	36,878	1.646	2.77	738	448
112	160,942	1.83	2.683	3.219	1.759
114	84,132	1.584	1.066	1.683	1.062
116	37,433	1.592	1.484	749	0.47
117	49,462	1.554	1.014	989	636
119	6,406	1.254	-348	128	102
121	105,590	1.655	1.585	2.112	1.276
122	5,234	1.191	-373	105	88
123	62,585	1.742	3.158	1.252	718
124	18,578	1.451	-13.273	372	256
127	65,155	1.737	2.858	1.303	0.75
128	73,555	1.553	1.024	1.471	948

B. Variant Calling Steps Codes

For VCF files, the following tools with corresponding scripts have been used:

1. cutadapt: To trim raw reads.
   Version: 2.6

   Script: $ cutadapt -q 20 -m 20 -a CTGTCTCTTATA -A CTGTCTCTTATA -o forward_trimmed.fastq -p reverse_trimmed.fastq forward_raw.fastq reverse_raw.fastq

2. bwa: To generate reference indexes and alignment.
   Version: bwa-0.7.17-r1188
   Reference genome: Ensembl, Homo_sapiens.GRCh38.dna.toplevel.fa
   Indexes were generated by `bwa index`:
   Script: bwa index Homo_sapiens.GRCh38.dna.toplevel.fa
   Alignment was made by `bwa mem`:
   Script: $ bwa mem Homo_sapiens.GRCh38.dna.toplevel.fa forward_trimmed.fastq reverse_trimmed.fastq > alignment.sam

3. samtools: To compress and sort sam files.
   Version: samtools 1.19.2
   BAM generation: samtools view -b -o alignment.bam alignment.sam
   BAM compression and sorting: samtools sort -l 9 -o alignment_sorted.bam alignment.bam
   BAI generation: samtools index -b alignment_sorted.bam alignment_sorted.bam.bai

4. freeBayes: To generate VCFs.
   Version: v1.3.7
   Script: freebayes -f Homo_sapiens.GRCh38.dna.toplevel.fa alignment_sorted.bam > variants.vcf

From created VCF files, following python scripts have been used to create meaningful, human-readable tables:

1.
```
import os
import pandas as pd


def extract_variant_status_for_patient(patient_number, expanded_regions, target_SNVs):
    vcf_filename = f'{patient_number}.vcf'
    if not os.path.exists(vcf_filename):
        print(f"File not found: {vcf_filename}")
        return {}


def get_patient_numbers_from_vcf(directory='.'):
    patient_numbers = []
    for filename in os.listdir(directory):
```

```
        if filename.endswith(".vcf"):
            patient_number = filename.split(".")[0]
            try:
                patient_number = int(patient_number)
                patient_numbers.append(patient_number)
            except ValueError:
                print(f"Invalid filename format: {filename}")
    return patient_numbers


patient_numbers = get_patient_numbers_from_vcf()

target_variants = [('17', 44229986),
                   ('21', 41481156), ('21', 41473715),
                   ('21', 41471515), ('21', 41464259),
                   ('21', 41491393), ('21', 41473447),
                   ('21', 33287378), ('19', 4717660),
                   ('X', 15621438), ('6', 32623820),
                   ('21', 33230000), ('2', 162073469),
                   ('11', 34482745), ('2', 60480453),
                   ('21', 41508407), ('21', 41508389),
                   ('21', 41508379), ('12', 132489231)]

window_size = 10

target_SNVs = {'rs2532300', 'rs4290734', 'rs2298661', 'rs34624090', 'rs463727',
               'rs35899679', 'rs2298659', 'rs8178521', 'rs12610495',
               'rs1548474', 'rs9271609', 'rs17860115', 'rs17574',
               'rs61882275', 'rs1123573', 'rs61299115', 'rs11088551', 'rs4303794', 'rs56106917'}

windows = [(chrom, max(0, start - window_size), start + window_size) for chrom, start in
target_variants]
# Here we create 20 bp windows for future work ease

def process_vcf(file_name, windows):
    with open(file_name, 'r') as file:
        patient_id = file_name.split('.')[0]
        for line in file:
            if line.startswith('#'):
                continue  # Skip header lines
            parts = line.strip().split('\t')
            chrom, pos = parts[0], int(parts[1])
            for window in windows:
                for item in parts:
                    if item.startswith("0/0"):
                        if chrom == window[0] and window[1] <= pos <= window[2]:
                            yield {
                                'Patient': patient_id,
                                'Chromosome': chrom,
                                'Position': pos,
                                'Reference': parts[3],
                                'Alternative': parts[4],
                                'Situation': 'HOM_REF'
                            }
```

```
                elif item.startswith("0/1"):
                    if chrom == window[0] and window[1] <= pos <= window[2]:
                        yield {
                            'Patient': patient_id,
                            'Chromosome': chrom,
                            'Position': pos,
                            'Reference': parts[3],
                            'Alternative': parts[4],
                            'Situation': 'HET'
                        }
                elif item.startswith("1/1"):
                    if chrom == window[0] and window[1] <= pos <= window[2]:
                        yield {
                            'Patient': patient_id,
                            'Chromosome': chrom,
                            'Position': pos,
                            'Reference': parts[3],
                            'Alternative': parts[4],
                            'Situation': 'HOM_ALT'
                        }


# List all VCF files in the current directory
vcf_files = [f for f in os.listdir('.') if f.endswith('.vcf')]

# Process each VCF file and collect the results
results = []
for file in vcf_files:
    results.extend(process_vcf(file, windows))

# Convert the results to a DataFrame
df = pd.DataFrame(results)

csv_filename = "output.csv"
df.to_csv(csv_filename, index=False)  # index=False to not write row indices
```

2.

```
import pandas as pd

data = pd.read_csv("output.csv")

locations = pd.read_csv("locations.csv")

matched_data = pd.DataFrame(columns=data.columns.tolist() + ['UniqueID'])

# Iterating through each row in 'data'
for index, row in data.iterrows():
    # Finding matching rows in 'locations'
    matched_rows = locations[(locations.iloc[:, 0] == row.iloc[1]) & (locations.iloc[:, 1] ==
row.iloc[2])]

    # If match is found, add the unique identifier and append the row to 'matched_data'
    if not matched_rows.empty:
        # Assuming the unique identifier is in the third column of 'locations'
        unique_id = matched_rows.iloc[0, 2]
```

```
        new_row = row.tolist() + [unique_id]
        matched_data = pd.concat([matched_data, pd.DataFrame([new_row],
columns=matched_data.columns)],
                                 ignore_index=True)

# Drop 'Chromosome' and 'Position' columns from the 'matched_data' DataFrame
matched_data = matched_data.drop(['Chromosome', 'Position'], axis=1)


def combine_columns(row):
    return row["Situation"]


matched_data['Combined'] = matched_data.apply(combine_columns, axis=1)

matched_data = matched_data.drop(['Reference', 'Alternative'], axis=1)
matched_data.drop_duplicates(inplace=True)
matched_data.reset_index(drop=True, inplace=True)

all_patients = matched_data['Patient'].unique()
all_unique_ids = matched_data['UniqueID'].unique()


# Creating a new DataFrame with all combinations of Patient and UniqueID
full_combinations = pd.MultiIndex.from_product([all_patients, all_unique_ids],
names=["Patient", "UniqueID"]).to_frame(index=False)

# Merging with the original DataFrame
merged_data = full_combinations.merge(matched_data, on=['Patient', 'UniqueID'],
how='left')

# Filling missing values in 'Combined' column with 'Wild'
merged_data['Combined'].fillna('Wild', inplace=True)


# Pivoting the DataFrame
reshaped_data = merged_data.pivot(index='UniqueID', columns='Patient',
values='Combined')

# Resetting the index to make UniqueID a column
reshaped_data.reset_index(inplace=True)


csv_filename = "merge.csv"
reshaped_data.to_csv(csv_filename, index=False)  # index = False to not write row indices
```

In this step a file “complete_table.csv” has been created by utilizing excel functions to omit readings with not enough read depths (10 in this study). Therefore complete_table.csv is a copy of merge.csv where information is labeled as NA if 10 read depth threshold is not fulfilled.

3.

```
import pandas as pd
import pandas as pd
import numpy as np
```

```
from scipy.stats import norm

master_freq = pd.read_csv("Master_freq.csv")
variants = pd.read_csv("complete_table2.csv")


def one_proportion_z_test(p_hat, p0, n):
    if pd.isna(p0):  # Handle missing known values
        return "NA"
    std_error = np.sqrt(p0 * (1 - p0) / n)
    if std_error == 0:
        return "Not calculable"  # Return a message indicating the test couldn't be performed
due to zero std error
    z_score = (p_hat - p0) / std_error
    p_value = 2 * (1 - norm.cdf(abs(z_score)))  # Two-tailed test
    return p_value


# Define a function to calculate frequencies for each row
def calculate_frequencies(rw):
    # Count occurrences of each value excluding NaN
    value_counts = rw.value_counts()
    WIL_freq = value_counts.get("WIL", 0) / rw.count()
    HET_freq = value_counts.get("HET", 0) / rw.count()
    HOM_freq = value_counts.get("HOM", 0) / rw.count()

    # Calculate frequencies for "WIL", "HET", "HOM" if they exist in the row, else 0
    frequencies = {
        "index": rw[0],
        "WIL_freq": WIL_freq,
        "HET_freq": HET_freq,
        "HOM_freq": HOM_freq,
        "Allele_Freq": HOM_freq + (HET_freq / 2),
        "Count": rw.count()
    }
    return pd.Series(frequencies)


df = variants.apply(calculate_frequencies, axis=1)


# Iterate through df rows
for index, row in df.iterrows():
    # Find the corresponding row in master_freq (assumes matching index or another
method of identification)
    master_row = master_freq.loc[index]
    # Perform Z-test for EU, comparing against 'Allele_Freq' from df
    p_value_EU = one_proportion_z_test(row['Allele_Freq'], master_row['EU'],
row['Count'])
    df.at[index, 'p_value_EU'] = p_value_EU


# Now df includes p-values for comparisons to both EU frequencies from master_freq
```

```
# Create a new DataFrame with only the required columns
result_table = df[['index', 'p_value_EU']]

print("z_test results for allele frequencies compared to EU and TR allele frequencies")
# Print the new DataFrame
print(result_table)

print("_____")
print("Starting bootstrap algorithm")


def bootstrap_CI(row, n=100, k=1000, confidence_levels=[0.95, 0.99]):
    allele_freq = row['Allele_Freq']  # Check the allele frequency

    # Create k populations with n simulated alleles each, based on allele frequency
    simulated_data = [np.random.binomial(n=1, p=allele_freq, size=n) for _ in range(k)]

    # Calculate the mean frequency of mutated alleles for each simulation
    simulated_means = [np.mean(simulation) for simulation in simulated_data]

    # Calculate the overall mean and standard deviation for the bootstrap samples
    overall_mean = np.mean(simulated_means)
    overall_std = np.std(simulated_means)

    # Calculate confidence intervals for each confidence level
    ci_results = {}
    for cl in confidence_levels:
        lower_bound = np.percentile(simulated_means, 100 * ((1 - cl) / 2))
        upper_bound = np.percentile(simulated_means, 100 * (1 - (1 - cl) / 2))
        ci_results[f'{int(cl * 100)}%_CI'] = (lower_bound, upper_bound)

    return overall_mean, overall_std, ci_results


# Iterate over rows in df
for index, row in df.iterrows():
    overall_mean, overall_std, ci_results = bootstrap_CI(row)

    # Store the mean and standard deviation results
    df.at[index, 'Bootstrap_Mean'] = overall_mean
    df.at[index, 'Bootstrap_Std'] = overall_std

    # Store the CI results
    df.at[index, '95%_CI_Lower'], df.at[index, '95%_CI_Upper'] = ci_results['95%_CI']

# Create a copy of df to ensure it's a separate DataFrame
df = df.copy()

# Set 'rs_code' as the index in a copy of master_freq for efficient mapping
master_freq_indexed = master_freq.set_index('rs_code')

# Map 'EU' frequencies to df using the appropriate column
df['EU'] = df['index'].map(master_freq_indexed['EU'])
```

```
# Check the first few rows to verify
print(df.head())


csv_name = "bootstrap_results.csv"
df.to_csv(csv_name, index=False)
print(f"Bootstrap results have been saved to {csv_name}")


print("_____")
print("Starting comparisons according to bootstrap")

for index, row in df.iterrows():
    eu_freq = master_freq.at[index, 'EU']
    mean = row['Bootstrap_Mean']
    std = row['Bootstrap_Std']
    ci_95_lower, ci_95_upper = row['95%_CI_Lower'], row['95%_CI_Upper']

    # Confidence Interval Overlap
    ci_overlap_eu = 'Yes' if ci_95_lower <= eu_freq <= ci_95_upper else 'No'
    z_score_sig = 2  # Change Z_score significance threshold if needed

# Handle divide by zero in standard deviation
    if std == 0:

        z_score_eu = 0 if mean == eu_freq else np.inf
        significance_eu = 'No' if mean == eu_freq else 'Yes'
    else:
        # Distance from Mean
        z_score_eu = abs(eu_freq - mean) / std
        significance_eu = 'Yes' if z_score_eu > z_score_sig else 'No'


    # Store results in DataFrame
    df.at[index, 'CI_Overlap_EU'] = ci_overlap_eu
    df.at[index, 'EU_Significantly_Different'] = significance_eu


csv_name = "bootstrap_comparison_results.csv"
df.to_csv(csv_name, index=False)
print(f"Bootstrap results have been saved to {csv_name}")
```

C. Test and Comparison Results

Table 6 Comprehensive results table

index	WIL freq	HET freq	HOM freq	Allele Freq	Count	p_value EU	Bootstrap Mean	Bootstrap Std	95%_CI Lower	95%_CI Upper	EU	CI Overlap
rs11088551	0.333	0.508	0.150	0.404	120	0.692	0.405	0.048	0.310	0.500	0.422	Yes
rs1123573	0.423	0.454	0.113	0.340	97	0.365	0.341	0.049	0.250	0.440	0.385	Yes
rs12610495	0.632	0.333	0.026	0.192	117	0.061	0.193	0.039	0.120	0.270	0.269	Yes
rs1548474	0.609	0.173	0.209	0.295	110	0.900	0.295	0.046	0.210	0.380	0.29	Yes
rs17574	0.491	0.430	0.070	0.285	114	0.224	0.284	0.045	0.200	0.370	0.339	Yes
rs17860115	0.273	0.121	0.576	0.636	33	0.000	0.633	0.048	0.540	0.730	0.324	No
rs2298659	0.720	0.254	0.017	0.144	118	0.025	0.144	0.035	0.080	0.220	0.231	No
rs2298661	0.718	0.256	0.017	0.145	117	0.043	0.144	0.033	0.080	0.210	0.223	No
rs35899679	0.397	0.431	0.164	0.379	116	0.061	0.379	0.049	0.280	0.470	0.466	Yes
rs4290734	0.959	0.000	0.031	0.031	97	0.000	0.030	0.017	0.000	0.070	0.489	No
rs4303794	0.333	0.508	0.150	0.404	120	0.692	0.406	0.049	0.320	0.500	0.422	Yes
rs463727	0.359	0.410	0.222	0.427	117	0.479	0.427	0.050	0.330	0.530	0.46	Yes
rs61882275	0.313	0.527	0.152	0.415	112	0.368	0.415	0.050	0.320	0.510	0.374	Yes
rs8178521	0.492	0.408	0.092	0.296	120	0.491	0.298	0.045	0.210	0.390	0.268	Yes
rs9271609	0.605	0.382	0.000	0.191	76	0.032	0.192	0.040	0.120	0.270	0.304	No
rs2532300	0.991	0.000	0.000	0	106	0.000	0	0	0	0	0.218	No
rs34624090	0.991	0.000	0.000	0	116	0	0	0	0	0	0.449	No
rs61299115	0.992	0.000	0.000	0	120	0	0	0	0	0	0.421	No
rs56106917	0.988	0.000	0.000	0	84	0	0	0	0	0	0.493	No

D. Read Depth Results

Table 7 Read depth mean and median values per location. Results below threshold indicates how many of the patients' results dropped below 10, therefore omitted form results.

Location	Mean	Median	Results below threshold (10)
rs2532300	1306.275	1079	3
rs4290734	1732.483333	1131	16
rs2298661	11156.34167	11967	4
rs34624090	6900.483333	6030	2
rs463727	35219.68333	30999	3
rs35899679	34687.5	21208.5	4
rs2298659	5379.283333	4057	2
rs8178521	17201.71667	15698.5	1
rs12610495	51413.56667	43899.5	4
rs1548474	5471.875	4921	7
rs9271609	785.6583333	180.5	16
rs17860115	161.4916667	26.5	32
rs17574	10563.25833	9356.5	7
rs61882275	15532.125	13576.5	7
rs1123573	421.4333333	271.5	3
rs61299115	26408.39167	19804	1
rs11088551	39804.25	37232	1
rs4303794	38028.94167	35662	1
rs56106917	343.8	180	3

E. Primer Assays

Table 8 Designed primer assays

SNV	Forward Primer	Reverse Primer	Sequencing Primer
rs61882275	CACCCCACCCTGATGAGAATAA	CGTCTTGCTTTCTACCCTTCATA	CTAGGAAGGCCCTGC
rs61299115	GTTCCCTGCCTGGCTCAAC	GGCTCACCCAGGACTCCA	CCCTCTCCGGTCCGC
rs56106917	GGCATCTGGCAAATACGC	TGGTGGGGTTTTCTGAGC	TTTTCTGAGCCCAGG
rs35899679	TTTTTTGAGACAGGGTCTTGC	GCAGCCGAGTTATGAGAACATC	GGGCCCAAGTGATCC
rs34624090	GGCAACACTGGTTGTCACAGCT	AAAAGGCTAGAACAGGGCTTTCC	CCCTACAA
			CAGAGAACCA
rs17860115	CACCCGCACTAAAGACGCTTC	GCTCGGGGAACTGTTCGG	CGGCGGGATTCCTAC
rs12610495	ATTATCTTGCCTCTGATCACCACT	AATCCATCTCTGCAGCTGTGTA	GCAAAAATTCGATCCC
rs11088551	GTTCCCTGCCTGGCTCAAC	GGCTCACCCAGGACTCCA	CGGCCCCAGCGGCCA
rs9271609	TTTTCATCACCTCCAAAGGAGACC	GGGCCTAAGCTCCTTCTGCATA	CCTCCAAAGGAGACC
rs8178521	GGCAGCCTTGGGTTTTTCC	CGAGGCTGTGATAGTGAGCTATGA	TGGTTTTGGTTTTTGA
rs4303794	GCTGCGGAGGGACCCATA	CGCCTACAGGAGCTCGTGA	GGGCAGGCTGGCCGC
rs4290734	AAATGTTCACTGCAACCCTCTTA	TTGCAGCCTGTGTCTGAATTT	ACTGCAACCCTCTTAAA
rs2532300	GGGCTACTTGGATGAAGGTATAAA	GGGGGGGATTTCAGAGATT	GCTCTTTTCAAACCCTT
rs2298661	ACTGCTGGGGATGGACTTAGG	GAGTCCCCAGGGCCTTGTTA	CCACTTCCCTGAACAC

SNV	Forward Primer	Reverse Primer	Sequencing Primer
rs2298659	CCCCTGGCATACTTTTCC	AACTCAAGCCGCCAGAGC	GAGCAGGATTGTGGG
rs1548474	AAAATCTGGATTTGTGGCAGAAG	ATCAGGCCAAATCACAGTAACAT	TTGTCAGTGCTTAAGGAA
rs1123573	GATTCAAGCGCTGTGCAGT	CCTTCCAGCCAGGTCATTTAGAAT	AACTTTTCTTCCTGTTGG
rs463727	TGGGTCTTTCCTGTGCTGTTTTT	GGTCCCTCAAATGACTCCTTCTTA	CCATGATTGTGAGGC
rs17574	CGCGGTCTCCCTCTTCTAAC	CACGGTGATGATGGTGACAAG	ACACCGTGGAAGGTT