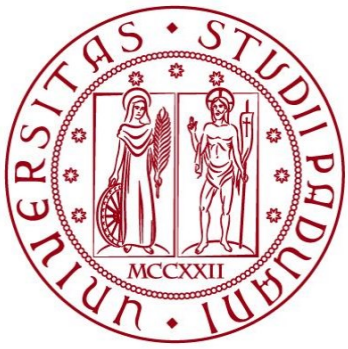
**UNIVERSITÀ DEGLI STUDI DI PADOVA**

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology




**TESI DI LAUREA**

# Association study of rare variants in human sepsis

**Relatore: Prof. Giovanni Vazza**
**Dipartimento di biologia**

**Correlatore: Prof. Henri-Jean Garchon**
**Université Paris Saclay / UMR1173**

**Laureando: Smail Syphax Zeggane**

**ANNO ACCADEMICO 2023/2024**

# Table of contents

**ABSTRACT** ..... **3**

**INTRODUCTION** ..... 4

**MATERIAL & METHODS** ..... 6

**RESULTS** ..... 8

**CONCLUSION** ..... 18

**DISCUSSION** ..... 19

**Abstract**

Sepsis is a highly lethal syndrome resulting from abnormal immune and metabolic responses to infection, thereby compromising host homeostasis. Previous genome-wide association studies (GWAS) have identified several genetic loci influencing susceptibility to sepsis disease. However, these loci have shown limited effects and account for only a small portion of the genetic component underlying sepsis susceptibility. In this study, we aimed to uncover rare protein-coding gene (PCG) variants associated with sepsis outcomes. We examined Whole Genome Sequencing (WGS) data of 701 patients with sepsis enrolled in the RHU RECORDS project. After accounting for multiple testing correction, our analysis revealed no detectable effect sizes at our current sample size. Nonetheless, we revealed a cluster of variants situated within the genes TEXT13B, FAM83A, NEDT19, and SLC9A3R1, collectively exhibiting a strong association with sepsis mortality. This knowledge is expected to provide novel insights into potential drug targets for sepsis, risk stratification, and therapeutic response. Characterizing genetic variants associated with sepsis outcomes is crucial to identifying high-risk patients who may benefit from more personalized interventions and individually targeted therapies.

**Intro:**

Throughout history, infectious diseases have been a significant threat to human health[1]. In response to the tissue damage triggered by pathogens, the body initiates an inflammatory response to restore homeostasis. This inflammatory response is complex and can be divided into three distinct phases: an initial alarm phase, where the release of inflammatory mediators is triggered by signals of danger; a subsequent mobilization phase, where leukocytes migrate to the site of injury; and finally, a resolution phase, where cellular debris are moved from the infected tissue. The effective completion and resolution of the inflammatory response is critical for proper healing[2].

Sepsis is a life-threatening condition that arises in response to infection, following the definition of the 3rd International Consensus Definition for Sepsis and Septic Shock[3]. While gram-positive or gram-negative bacteria are typically the causative agents, sepsis can also result from infection with fungi, viruses, and parasites[4]. It results from a systemic activation of the innate immune system that, in combination with intense complement activation, turns what would normally be a physiological response to infection into an excessive inflammatory response. Consequently, tissue damage, cellular compromise, and molecular dysregulation occur, initiating organ dysfunction, and possibly leading to multi-organ failure and death[5]**.**

Clinically, organ dysfunction is assessed by an increase in the Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score of 2 points or more, which is associated with an in-hospital mortality greater than 10%**.** In 2017, the World Health Organization (WHO) member states declared that improvement of sepsis prevention, recognition, and treatment was a global health priority and the most urgent unmet medical need of our times[6] and called again for global action in 2020. Sepsis was the

most expensive condition in US hospitals in 2011, comprising 6.2% of aggregate costs for all hospitalizations, a sum of 23.7 billion USD annually[7]. The pooled incidence of hospital-treated sepsis cases is 189 per 100,000 person-years. In intensive care units (ICUs), the mortality rate prior to hospital discharge for patients with sepsis was 41.9%[8]. The incidence of sepsis in most Low- or Middle-Income Countries (LMICs) is still unknown, emphasizing the pressing need for enhanced epidemiological surveillance.

Furthermore, prolonged sepsis can lead to an immunocompromised state in which the host immunologic response evolves from a hyper-inflammatory state to an anti-inflammatory state[4]. Thus, survivors of sepsis have an increased risk of death for up to five years after the acute illness[9], and their quality of life is significantly impaired[10,11]. Hence, researchers proposed therapeutic interventions that curb hyperinflammation, promote anabolism, and bolster immune function[12]. One example is the discovery of synthetic glucocorticoids in 1948 that revolutionized the treatment of immune-related disorders. Their effectiveness was recognized by the award of the Nobel Prize in Physiology or Medicine in 1950 to Philip S. Hench, Edward Kendall, and Tadeus Reichstein. Due to their potent immunosuppressive effects, glucocorticoids (GCs) remain the mainstay in the treatment of numerous inflammatory and autoimmune pathologies. Likewise, since the mid-twentieth century, corticosteroids have been used as adjuvant therapy for sepsis to counteract various sepsis characteristics such as excessive inflammation, vascular defects, and hypoglycemia[13]. Sepsis is however complicated by dysfunction of the HPA (hypothalamic-pituitary-adrenal) axis, caused by critical-illness-related corticosteroid insuficiency and GC resistance, a well-recognized manifestation in sepsis that may hinder their eficacy[4].

Thus, several clinical trials have evaluated the safety and eficacy of glucocorticoids in sepsis[14,15,16,17]. However, such interventions have often been applied

unselectively to heterogenous patient groups without accounting for host genetic diversity on response to treatment[18], which is known to influence the host's immune response to microbial agents[19,20]. Characterizing genetic variants associated with sepsis outcomes is therefore expected to be crucial to identifying high-risk patients who may benefit from more personalized interventions and individually targeted therapies. This knowledge should also offer new insights into sepsis potential drug targets, risk stratification, and response to therapy. Genome-wide association studies (GWAS) have identified several genetic loci influencing susceptibility to infectious disease using a polygenic model of complex multifactorial traits[18,21,22,23,24,25]. However, these loci had small-size effects and contributed a small fraction of the genetic component of sepsis susceptibility. Indeed, the standard GWAS and imputation methods cannot test the effects of rare large-effect variants, as they are not commonly present in reference haplotypes and are not covered by genotyping arrays[26]. To overcome this gap, direct whole genome sequencing (WGS) or whole-exome sequencing (WES) are presently the technologies of choice to access rare variants.

In this study, we aim to identify rare protein-coding gene variants associated with sepsis outcomes. For this purpose, we analyzed Whole Genome Sequencing data of 701 patients with sepsis enrolled in the University-hospital research Rapid rEcognition of COrticosteRoiD resistant or sensitive Sepsis (RHU RECORDS) project.

## Material & methods:

### Contributing cohorts: Study design and participants

Detailed information regarding the design and implementation of the RECORDS TRIAL and RECORDS APROCCHS trials, including the trial protocol, amendments, and statistical analysis plan, has been previously documented[15,17]. The trial protocols, which encompassed genetic analyses, were approved by the Ethics Committee (Comité de Protection des Personnes, CPP) of Saint-Germain-en-Laye, France. Written informed consent was obtained from participants or their legally authorized representatives, and in certain cases, deferred written informed consent was obtained from patients. All authors were granted complete and independent access to the data, and they afirm the integrity, accuracy, and comprehensiveness of the data and analyses, as well as adherence to the study protocol.

Among the 1241 patients enrolled in the APROCCHS trial, Whole-Genome Sequencing (WGS) data were available for 261 patients. Collection of WGS data for the RECORDS TRIAL is still ongoing.

### Work environment

Data processing and variant calling procedures were executed using the Bash scripting language. Subsequently, variant functional annotation was performed using ANNOVAR within the Bash environment. For variant interpretation and statistical analysis, the R programming language (version 4.2.3) was employed.

### Principal Component Analysis

Genetic association analysis poses technical complexities due to the necessity of addressing potential confounding factors arising from disparities in population structure or relatedness, which can lead to an elevated risk of type 1 error (erroneously

rejecting the null hypothesis when it is true, resulting in false positives). To facilitate this, one approach is to employ principal component analysis, a mathematical technique used to summarize the primary sources of variance in the data. This method enables the modeling of ancestry-related differences and the consideration of divergent allele frequencies among distinct populations. (Know et al).

In this study, variants of autosomes with MAF above 10% were extracted from the 1000 genome project using BCFTOOLS[35] and intersected with variants from the RECORDS cohort. The resulting set of variants was used to compute 6 Principal Components by PLINK using samples from the 1000 genome projects. RECORDS samples were projected along the same Principal Components. To avoid bias in the analysis due to the different ethnicity, only patients of genetic European ancestry were retained for further analyses.

## Genetic scoring models and association analysis

We computed the cumulative occurrence of protein-coding variants across patient groups, using three distinct genetic models, namely additive, dominant, and recessive (Table 1).

**Table 1. Implementation of mathematical models scoring the occurrence of the effect allele**

**For a biallelic locus with 'A' for reference allele and 'a' for alternative allele, the given genotype score is :**

| Model | Genotype : | A/A | A/a | a/a |
|---|---|---|---|---|
| Additive | | 0 | 1 | 2 |
| Dominant | | 0 | 1 | 1 |
| Recessive | | 0 | 0 | 1 |

Subsequently, we conducted a comparative analysis to evaluate the allele frequencies and determine if they exhibit similarity or dissimilarity across patient groups for a given variant. This assessment involved performing a Fisher's exact test on the observation counts of individual alleles obtained through the utilization of the three distinct mathematical models.

**Results:**

**Filtering genetic variants**

In order to identify rare genetic variations (with a minor allele frequency [MAF] less than 1%) associated with sepsis outcomes, we conducted an analysis using whole genome sequencing data from 701 individuals across three distinct cohorts: APROCCHS (a randomized double-blind clinical trial conducted by **Annane et al.**), COHORT (an observational cohort study conducted during the COVID-19 crisis), and TRIAL (an ongoing randomized adaptive-Bayesian clinical trial conducted by **Fleuriet et al.).**

In total, we compiled a dataset comprising 59.8 million genetic variants (fig. 1a). Subsequently, we applied multiple filtering steps, including the removal of intergenic variants, of variants with a CADD score below 20 (Combined Annotation Dependent Depletion, a metric used to predict the deleteriousness or pathogenicity of genetic variants) and variants with a MAF > 1%. We also selected variants with "PASS" filter, (indicating high quality). This process resulted in a final set of 1.7 million variants, approximately 60% of which were found to be located within intronic regions (fig. 1b). To focus our analysis, we specifically examined protein-coding rare variants, which constituted approximately 3% (49040 variants) of the total (1.7 million). This choice was based on several factors, including the greater understanding of protein-coding

regions (than of noncoding ones), the relative ease of predicting their functional impact, and the convenience of experimental validation.

In addition, we evaluated the eficacy of the CADD score, a recent but presently widely used computational tool in the field of genomics research and clinical genetics for predicting the harmfulness or disease-causing potential of genetic variants. For this purpose, we examined the distribution of CADD scores among rare protein-coding variants, stratified according to their clinical significance groups, determined, and curated by expert geneticists, and reported in the ClinVar database. Our analysis revealed a significant disparity in CADD scores between variants known to be pathogenic or likely pathogenic, and those classified as benign, likely benign, or variants with uncertain significance (fig. 1c). This finding underscores the utility of the CADD score as a valuable tool for distinguishing between benign and potentially disease-causing variants. Based on the observed distribution, which indicated that most variants in the benign, likely benign, and uncertain significance groups possessed a CADD score below 30, we made the decision to exclude variants with a CADD score lower than 30. Consequently, our final dataset comprised 2686 rare exonic variants (fig.1a).

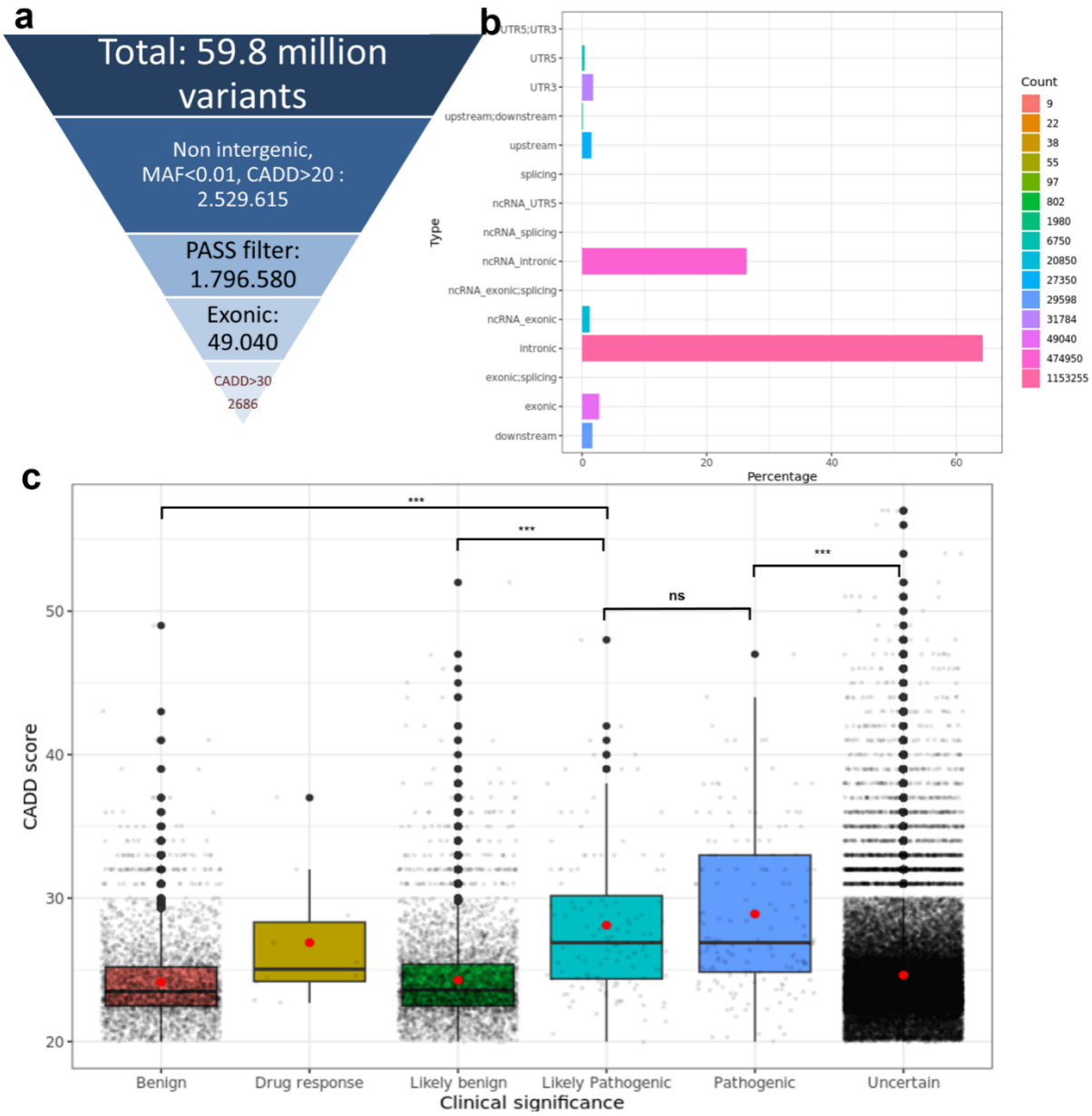



**Figure 1. Filtering processes of whole genome sequencing data.** a) Progression of variants number at each filtering procedure. b) Proportion and count of variants at each genomic location. c) CADD score distribution of rare protein-coding variants per clinical significance groups. Red dot represents the arithmetic mean. ANOVA (analysis of variance) was performed, followed by Tukey's multiple comparison of means and a non-parametric Kruskal-Wallis rank sum test. All statistical tests yielded p-values < 2e-

16. Significance levels used are as follows : $p \simeq 0$ (***), $p < 0.001$ (**), $p < 0.01$ (*), $p < 0.05$ (.), and (ns) for not significant.

**Population study**

To ensure a comprehensive analysis of our dataset, we investigated the population structure and composition by major ancestry groups. Initially, we conducted principal component analysis (PCA) on the genetic data of 701 patients, revealing the diverse ethnic composition within our population (fig. 2a). Additionally, we performed separate PCA analyses on patients belonging to the APROCCHS cohort, unveiling a predominant European genetic profile (fig. 2b). To mitigate the potential bias arising from population stratification, we focused our subsequent analysis exclusively on the 261 patients from the APROCCHS cohort (fig. 2c).

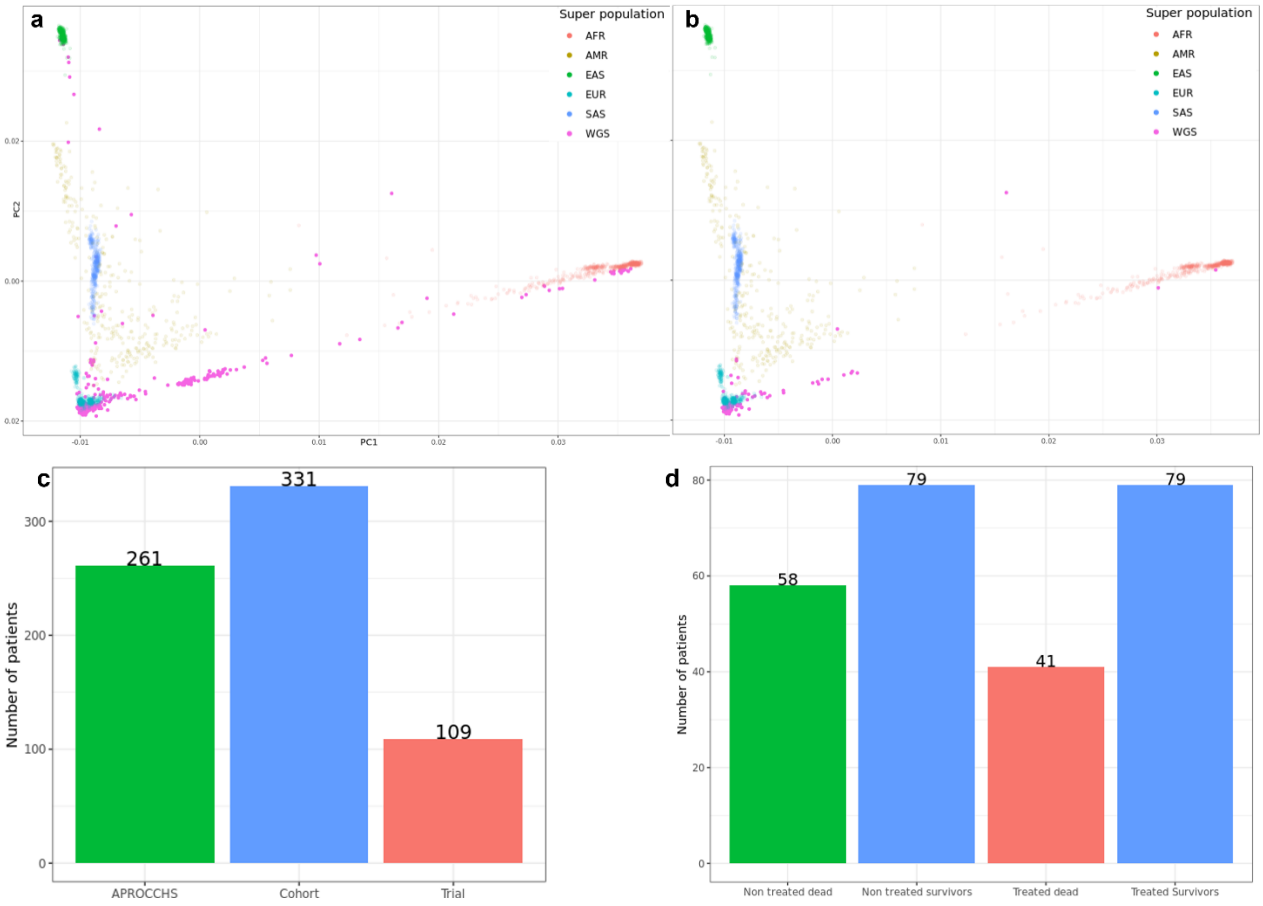



**Figure 2.** Overview of the contributing studies to the RHU RECORDS project and composition by major ancestry groups. The populations within the RECORDS samples were classified based on 1000 Genome project continental ancestry's data, including African (AFR), admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS), with WGS representing the whole genome sequencing data. The correlation observed between the WGS data, and the other populations indicates the alignment of the major ancestry groups represented in the WGS dataset. a) Principal Component Analysis (PCA) on samples derived from all contributing studies within RECORDS project. b) PCA on samples derived from APROCCHS study. c) Total number of patients included in each contributing cohort for the RECORDS project. d) Distribution of patient across groups in the APROCCHS cohort.

This decision was motivated by the fact that APROCCHS is a double-blind randomized clinical trial, while the TRIAL cohort's data are still being collected. Notably, patients from COHORT were independently analyzed, although caution is warranted due to the inherent risk of errors stemming from population stratification bias. Furthermore, for subsequent association analyses, we categorized patients of the APROCCHS cohort into four distinct groups based on their treatment and on their survival outcomes (fig. 2d).

**Allele occurrence count using genetic models**

To comprehensively analyze the rare protein-coding variants, we used three genetic models (additive, dominant, and recessive) to evaluate the cumulative occurrence of these variants across different patient groups (Table 1).

We employed a Venn diagram to investigate the presence of shared or specific rare variants within different patient groups (fig. 3a). Variants exclusive to specific groups were identified, but their impact and interplay remain inconclusive, as they were observed in only one or two patients, thereby limiting our ability to draw meaningful interpretations from these findings. Moreover, 37 variants were shared across all patient groups. Among them, we identified 4 highly recurrent rare PCG variants (protein-coding gene variants) independently of the scoring models (Table 2). The 'TAF1B' gene, which encodes a TATA-Box Binding Protein Associated Factor involved in transcription initiation by RNA polymerase I, has been implicated in various orofacial cleft disorders.

| Table 2. Most recurrent rare protein-coding gene variants in APROCCHS patients | | | | | | | |
|---|---|---|---|---|---|---|---|
| Gene | Variant | Effect | MAF | Treated Survived | Treated Deceased | Placebo Survived | Placebo Deceased |
| **Allele occurrence count using dominant genetic model[a]** | | | | | | | |
| TAF1B | rs400917 | Stop gain | 1E-4 | 79 | 41 | 79 | 58 |
| OR11G2 | rs200876108 | Stop gain | 5E-3 | 72 | 39 | 68 | 56 |
| KRT83* | rs2857667 | Stop gain | 6E-3 | 59 | 23 | 53 | 37 |
| HOMEZ | rs765691793 | Frameshift deletion | 1E-3 | 52 | 23 | 46 | 42 |

Effect, functional consequence of the variant. MAF, minor allele frequency.

[a]Individuals were considered to have 0 score of the effect allele if they were homozygous for the reference allele, 1 score of the effect allele if they were heterozygous, and 1 score if they were homozygous for the alternative allele.

*Variants curated benign or likely benign.

The 'OR11G2' gene corresponds to an olfactory receptor that facilitates interactions with odorant molecules in the nasal cavity, leading to the initiation of neuronal responses associated with smell perception. 'KRT83' gene belongs to the keratin gene family and contributes to the development of hair and nails. Lastly, the 'HOMEZ' variant encodes a homeobox and leucine zipper protein that is predicted to regulate transcription by RNA polymerase II. These variants could potentially serve as markers indicative of a predisposition to sepsis, although extensive analyses are required to test this hypothesis thoroughly.

## Genetic association study of APROCCHS cohort

To investigate the potential genetic associations with sepsis outcomes, we conducted a comparative analysis of allelic frequencies across different patient groups. Specifically, we performed Fisher's exact test to evaluate whether the allele frequencies varied significantly for specific genetic markers, using the observation counts obtained from three distinct genetic models.

Initially, our focus was on exploring genetic associations with interferon pathway genes, genes located in susceptibility loci identified through COVID-19 genome-wide association studies (GWAS), glucocorticoid signaling genes, and other immunologically relevant genes with potential therapeutic implications. Across all these genes, we were able to identify only one, 'TYK2', a tyrosine kinase gene previously curated in a COVID-19 GWAS. This gene was present in only two patients and did not exhibit a significant association with sepsis outcomes.

Furthermore, we investigated the potential relationship between genetic variants and the responsiveness to glucocorticoid treatment in sepsis patients by comparing allelic frequencies between treated patients who survived or died. We calculated p-values and odds ratios, indicating the probability of a given allele being protective or causative. Interestingly, we identified two genes ($p < 5E-02$) for both additive and dominant models, while no significant associations were observed for the recessive model (Table 3). The gene 'EPS8L2' is believed to play a role in membrane rufling and actin cytoskeleton remodeling, whereas 'SLC5A10' is a member of the sodium/glucose transporter family, frequently associated with histidine metabolism diseases.

Subsequently, we investigated whether our protein-coding rare variants were associated with patient survival or death in sepsis. Our analysis revealed four genes for both additive and dominant models (Table 3). Additionally, the 'OR11G2' variant mentioned above demonstrated significance for the recessive model with a p-value below 4E-02. 'TEX13B' is a gene predominantly expressed in the testis (limited information available). 'FAM84A' is involved in cell population proliferation and epidermal growth factor receptor signaling pathway. 'NUDT19' is predicted to enable magnesium ion binding activity. ‘NHERF1’ gene encodes a sodium/hydrogen exchanger regulatory cofactor. Diseases associated with NHERF1 include Nephrolithiasis (Kidney stones) and Hypophosphatemia (vitamin D-resistant rickets).

However, when we examined the collective effect of grouping these variants together, we observed a significant association with death (Odds ratio >18) in sepsis patients (p-value < 1E-07). This suggests that the association may involve multiple rare variants, each potentially exhibiting relatively high penetrance. **(Schork & al)**.

**Table 3. Associations between sepsis outcomes and PCG rare variants in APROCCHS patients (additive model[a])**

| Gene | Variant | Effect | Odds Ratio | P Value | MAF | CADD SCORE | Treated Survived | Treated Deceased |
|---|---|---|---|---|---|---|---|---|
| **Treated deceased vs treated surviving** | | | | | | | | |
| EPS8L2 | rs184903112 | Non-syn SNV | 2.1 | 3.9E-2 | 1.4E-3 | 33 | 0 | 3 |
| SLC5A10 | rs61741107 | Non-syn SNV | 2.1 | 3.9E-2 | 4.1E-3 | 33 | 0 | 3 |
| Gene | Variant | Effect | Odds Ratio | P Value | MAF | CADD SCORE | Survived | Deceased |
| **Deceased vs surviving** | | | | | | | | |
| TEX13B[b] | rs41304466 | Stop gain | Inf | 2.2E-2 | 6E-3 | 33 | 0 | 4 |
| FAM83A | rs148011353 | Stop gain | 11.54 | 6.3E-3 | 3.8E-3 | 35 | 1 | 7 |
| NUDT19 | rs148011353 | Non-syn SNV | 9.84 | 1.5E-2 | 9.4E-3 | 32 | 1 | 6 |
| NHERF1 | rs41282065 | Non-syn SNV | Inf | 2.2E-2 | 2.6E-3 | 32 | 0 | 4 |
| **Patients carrying at least 1 variant associated with mortality (variants pooled)** | | | | | | | | |
| Variants group[c] | | - | 18.62 | 1E-07 | - | | 2 | 21 |

Effect functional consequence of the variant. P value, before FDR multiple testing correction.

[a]Individuals were considered to have 0 score of the effect allele if they were homozygous for the reference allele, 1 score of the effect allele if they were heterozygous, and 2 score if they were homozygous for the alternative allele.

[b]TEX13B is located on the X chromosome. Hemizygous males are included in the N of individuals with two copies of the effect allele.

[c]Group of variants associated with mortality, located in genes TEXT13B, FAM83A, NEDT19, SLC9A3R1.

*Variants curated benign or likely benign.

**Contribution of Sequential Organ Failure Assessment (SOFA) score in Sepsis**

Furthermore, we aimed to explore the association between the Sequential Organ Failure Assessment (SOFA) score at inclusion and mortality in sepsis patients. Our analysis revealed that patients who survived the acute illness had significantly lower SOFA scores compared to those who succumbed to the condition (fig. 3b). Considering this observation, we investigated whether the treatment modality influenced the trajectory of survival outcomes among patients presenting with a high SOFA score.

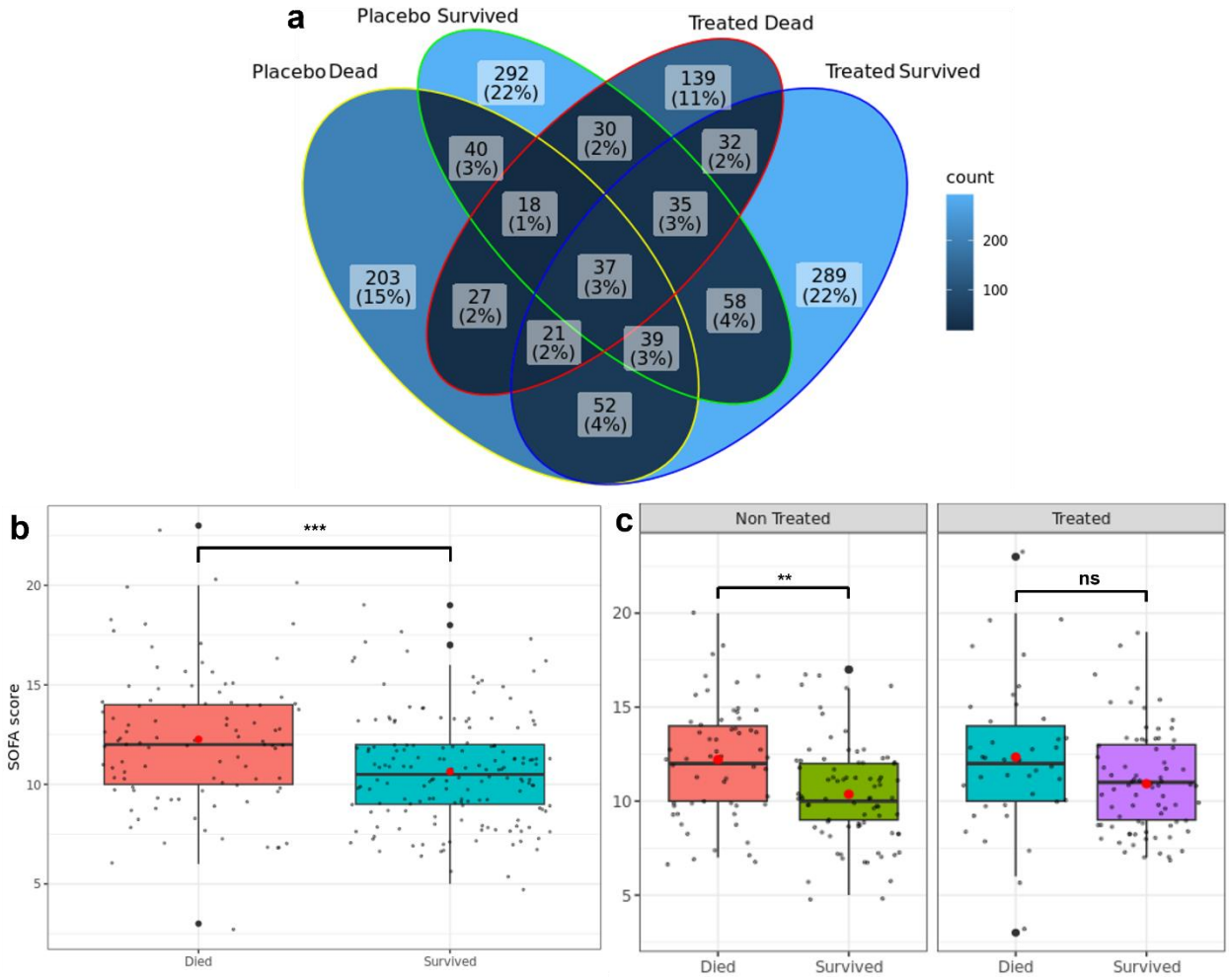



**Figure 3.** Comprehensive evaluation of the prognostic utility of the Sequential Organ Failure Assessment (SOFA) score in predicting sepsis outcomes. a) Comparative analysis of variant distribution among distinct patient groups through a Venn

diagram, highlighting the presence of specific and shared variants within each group. b) SOFA score distribution across sepsis survival and mortality outcomes. c) SOFA score distribution among treated and non-treated patients by mortality and survival outcomes. b-c) ANOVA (analysis of variance) was performed, followed by Tukey's multiple comparison of means and a non-parametric Kruskal-Wallis rank sum test. All statistical tests yielded p-values < 5e-2. Significance levels used are as follows: $p \simeq 0$ (***), $p < 0.001$ (**), $p < 0.01$ (*), $p < 0.05$ (.), and (ns) for not significant. Red dot represents the arithmetic mean.

Interestingly, we observed that among non-treated patients, those who died had higher SOFA scores compared to those who survived (fig. 3c). Conversely, among treated patients, the difference in SOFA scores between patients that survived and patients that died was not statistically significant. This suggests that the treatment may have influenced the severity of sepsis, as patients with initially high SOFA scores may have experienced improved survival outcomes following treatment, demonstrating a general trend of better survival among treated individuals.

**RECORDS COHORT analysis**

In our continued investigation, we conducted a comprehensive analysis of patients belonging to COHORT to identify potential associations between rare protein-coding variants and sepsis outcomes. As previously described, we assessed the collective presence of all rare protein-coding variants within various patient groups, employing three distinct scoring models: additive, dominant, and recessive (Table 1, mat & methods).

Interestingly, most recurrent rare variants across all patients were consistent across the three scoring models. Notably, these findings align with the variants identified in the

APROCCHS cohort (Table 4). These findings provide supporting evidence for the hypothesis that these variants may serve as markers indicative of susceptibility to sepsis. Nevertheless, a more robust investigation involving control vs. case analyses would be essential to further elucidate and substantiate this hypothesis.

**Table 4. Most recurrent rare protein-coding gene variants in RECORDS COHORT patients**

| Gene | Variant | Effect | MAF | Treated Survived | Treated Deceased | Placebo Survived | Placebo Deceased |
|---|---|---|---|---|---|---|---|
| **Allele occurrence count using dominant scoring model**[a] | | | | | | | |
| TAF1B | rs400917 | Stop gain | 1E-4 | 141 | 59 | 39 | 9 |
| OR11G2 | rs200876108 | Stop gain | 5E-3 | 123 | 49 | 37 | 9 |
| KRT83* | rs2857667 | Stop gain | 6E-3 | 87 | 44 | 24 | 4 |
| HOMEZ | rs765691793 | Frameshift deletion | 1E-3 | 86 | 38 | 22 | 6 |

Effect, functional consequence of the variant. MAF, minor allele frequency.

[a]Individuals were considered to have 0 score of the effect allele if they were homozygous for the reference allele, 1 score of the effect allele if they were heterozygous, and 1 score if they were homozygous for the alternative allele.

*Variants curated benign or likely benign.

In our investigation, we identified a single variant within the gene 'KRT83' that displayed a potential association with a favorable response to glucocorticoids among sepsis patients (Table 5). 'KRT83' is prominently linked to Monilethrix, a condition characterized by impaired hair growth. Three additional variants were implicated in a potentially unfavorable response to glucocorticoids in sepsis patients. Among them, 'KIAA1522' represents an uncharacterized gene, while the other two variants are curated benign based on existing literature.

Furthermore, we discovered five variants that exhibited a potential association with mortality in sepsis patients. One of these variants corresponds to 'KIAA1522,' which was discussed in the preceding paragraph. The remaining variants are curated benign.

However, it is important to exercise caution when interpreting these results due to population stratification and the heterogeneity of patient care.

Moreover, we conducted an independent analysis to explore the relationship between survival and COVID-19 infection status. Specifically, we focused on patients who were either positive or negative for COVID-19. Among COVID-19-negative patients, we observed a noteworthy association between mortality in sepsis patients and a variant located within the gene 'NLGN2', member of the neuronal cell surface protein family.

In the case of COVID-19-positive patients, we identified three variants that exhibited an association with mortality. The gene 'TBC1D2,' which has been curated as benign, is known to be involved in the positive regulation of GTPase activity. 'PPP1R15A' is part of a group of genes that display increased transcript levels under conditions of stressful growth arrest and in response to DNA-damaging agents, with its protein response being linked to apoptosis. 'TMEM177' functions as an integral component of the mitochondrial inner membrane and plays a role in mitochondrial complex IV assembly. However, it is important to note that after applying the false discovery rate (FDR) multiple testing correction, these findings did not reach statistical significance.

**Table 5. Associations between sepsis outcomes and PCG rare variants in COHORT patients (additive model[a])**

| Gene | Variant | Effect | Odds ratio | P value | MAF | CADD SCORE | Treated survived | Treated deceased |
|---|---|---|---|---|---|---|---|---|
| **Treated deceased vs treated surviving** | | | | | | | | |
| KRT83* | rs2857667 | Stop gain | 5E-2 | 1E-2 | 6.7E-3 | 36 | 110 | 59 |
| KIAA1522 | Unknown, MAF < 2E-4 | Frameshift insertion | 1.52 | 2E-2 | 2E-4 | 32 | 1 | 4 |
| TBC1D2* | rs137868712 | Non-syn SNV | 1.52 | 2E-2 | 3E-3 | 32 | 1 | 4 |
| DSC2* | rs200056085 | Frameshift insertion | 1.52 | 2E-2 | 8.7E-3 | 33 | 1 | 4 |
| Gene | Variant | Effect | Odds ratio | P value | MAF | CADD SCORE | Survived | Deceased |
| **Deceased vs surviving** | | | | | | | | |
| UNC93A* | rs145360877 | Stop gain | 7.5E-1 | 4E-2 | 2.7E-3 | 36 | 2 | 4 |
| KIAA1522 | Unknown | Frameshift insertion | 1.52 | 2E-2 | 2E-4 | 32 | 1 | 4 |
| GNA14* | rs138686336 | Non-syn SNV | 1.52 | 2E-2 | 2.8E-3 | 32 | 1 | 4 |
| DSC2* | rs200056085 | Frameshift insertion | 7.5E-1 | 4E-2 | 8.7E-3 | 33 | 2 | 4 |
| EPX* | rs35617692 | Stop gain | Inf | 5E-3 | 7.5E-3 | 34 | 0 | 4 |
| **COVID-19 negative deceased vs surviving** | | | | | | | | |
| NLGN2 | Unknown, MAF < 3E-04 | Frameshift deletion | 5.5 | 3E-2 | 3E-4 | 31 | 2 | 5 |
| **COVID-19 positive deceased vs surviving** | | | | | | | | |
| TMEM177 | rs145360877 | Stop gain | 9.71 | 4.5E-2 | 7.6E-3 | 40 | 1 | 3 |
| TBC1D2* | Unknown | Frameshift insertion | 9.71 | 4.5E-2 | 3E-3 | 32 | 1 | 3 |
| PPP1R15A | rs138686336 | Non-syn SNV | 9.71 | 4.5E-2 | 3.4E-3 | 32 | 1 | 3 |

Effect, functional consequence of the variant. P value, before FDR multiple testing correction.

[a]Individuals were considered to have 0 score of the effect allele if they were homozygous for the reference allele, 1 score of the effect allele if they were heterozygous, and 2 score if they were homozygous for the alternative allele.

* Variants curated benign or likely benign

## Conclusion

In summary, we explored the role of rare PCG variants on Sepsis outcomes based on whole genome sequence data, capturing genetic variation not assayed by array genotyping or imputation. After accounting for multiple testing correction, we did not identify any clear associations with rare protein-coding variants either exome wide or when specifically focusing on interferon pathway genes, genes located in susceptibility loci identified by COVID-19 GWAS[27,28,29,30,31,32], glucocorticoid signaling genes, and other additional genes of immunologic relevance and/or therapeutic potential. Our analysis indicated there are no detectable effect sizes at our current sample size. Nevertheless, our investigation unveiled a cluster of variants situated within the genes TEXT13B, FAM83A, NEDT19, and SLC9A3R1, which collectively demonstrate a robust association with sepsis mortality. These findings emphasize the potential contribution of genetic factors to Sepsis susceptibility and severity, offering valuable insights into disease pathogenesis and highlighting potential targets for therapeutic development or repurposing of existing drugs. Given the urgent need for effective Sepsis treatment strategies, further exploration of these genetic factors is warranted.

## Discussion

The genetic contribution to susceptibility of common complex diseases like Sepsis is still debated. The 'Common Disease, Common Variant (CDCV)' hypothesis argues for common variants with low penetrance (the probability that a carrier of the relevant variants will express the disease) as major contributors, while the 'Common Disease, Rare Variant (CDRV)' hypothesis argues for multiple rare variants, each with relatively high penetrance[33]. Rare mutations of variable penetrance cause single-gene inborn errors of immunity that lead to life-threatening infectious diseases[26,34]. Within the realm of scientific inquiry, numerous justifications support the proposition that the

combined influence of multiple rare variants, both intra- and inter-genic, exerts a substantial impact on the manifestation and prevalence of traits and diseases in the population.

First, the population dynamics, exemplified by the recent expansion of the human population, are posited to have engendered a substantial reservoir of segregating rare variants that possess functional significance, thus contributing to phenotypic diversity. Second the discovery of rare somatic mutations occurring independently in genes, which contribute to tumorigenesis, parallels the functional impact of inherited variants implicated in congenital. Third, the identification of multiple rare variants within a single gene associated disorders, including Cystic Fibrosis and BRCA1/BRCA2-associated breast cancer, indicates that rare variants may also influence common complex traits and diseases. Fourth, the identification of multiple functional variants within a specific gene, coupled with their associations with both in vitro and clinical phenotypes, implies that multiple rare variants could contribute to general clinical phenotypic expression. Finally, investigations employing targeted sequencing strategies focused on specific genes have provided empirical evidence of associations between collections of rare variants and specific phenotypes.

However, it is important to note that statistical analysis methods for examining the relationship between rare variants and phenotypes of interest are required. Unlike analyses involving common variations, association analyses involving rare variants are more challenging due to the low power to detect associations with a single rare variant, even in large sample sizes. Therefore, aggregation tests that evaluate the cumulative effects of multiple genetic variations within a gene or region have been proposed as a solution. These tests pool information from multiple variations at a specific locus, generating a burden score that is then tested for association with the trait or disease of interest. However, it should be acknowledged that these tests also

have limitations. "Burden" tests assume that all variations within a genetic unit have the same effect and direction, thus reducing power in the presence of neutral or opposite-effect variations.

Enhanced comprehension of the genetic architecture of diseases, along with a deeper understanding of the diverse forms and functionalities of DNA sequence variation, will inevitably influence the selection of appropriate statistical methodologies for conducting rare variant association studies.

**References:**

[1] Zhu, N. *et al.* (2020) 'A Novel Coronavirus from Patients with Pneumonia in China, 2019', *The New England Journal of Medicine*, 382(8), pp. 727–733. Available at: https://doi.org/10.1056/NEJMoa2001017.

[2] Cain, D.W. and Cidlowski, J.A. (2017) 'Immune regulation by glucocorticoids', *Nature reviews. Immunology*, 17(4), pp. 233–247. Available at: https://doi.org/10.1038/nri.2017.1.

[3] Singer, M. *et al.* (2016) 'The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)', *JAMA*, 315(8), pp. 801–810. Available at: https://doi.org/10.1001/jama.2016.0287.

[4] Vandewalle, J. and Libert, C. (2020) 'Glucocorticoids in Sepsis: To Be or Not to Be', *Frontiers in Immunology*, 11, p. 1318. Available at: https://doi.org/10.3389/fimmu.2020.01318.

[5] Rittirsch, D., Flierl, M.A. and Ward, P.A. (2008) 'Harmful molecular mechanisms in sepsis', *Nature Reviews. Immunology*, 8(10), pp. 776–787. Available at: https://doi.org/10.1038/nri2402.

[6] World Health Organization (2017) World Health Assembly 70, Resolution 70.7: improving the prevention, diagnosis, and clinical management of sepsis. Available at: http://apps.who.int/gb/ebwha/pdf_files/WHA70/A70_R7-en.pdf

[7] Torio CM, Moore BJ (2016) National inpatient hospital costs: the most expensive conditions by payer, 2013: statistical Brief #204Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Rockville (MD).

[8] Fleischmann-Struzek, C. *et al.* (2020) 'Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis', *Intensive Care Medicine*, 46(8), pp. 1552–1562. Available at: https://doi.org/10.1007/s00134-020-06151-x.

[9] Quartin, A.A. *et al.* (1997) 'Magnitude and duration of the effect of sepsis on survival. Department of Veterans Affairs Systemic Sepsis Cooperative Studies Group', *JAMA*, 277(13), pp. 1058–1063.

[10] Winters, B.D. *et al.* (2010) 'Long-term mortality and quality of life in sepsis: a systematic review', *Critical Care Medicine*, 38(5), pp. 1276–1283. Available at: https://doi.org/10.1097/CCM.0b013e3181d8cc1d.

[11] Storgaard, M. *et al.* (2013) 'Short- and long-term mortality in patients with community-acquired severe sepsis and septic shock', *Scandinavian Journal of Infectious Diseases*, 45(8), pp. 577–583. Available at: https://doi.org/10.3109/00365548.2013.786836.

[12] Delano, M.J. and Ward, P.A. (2016) 'The immune system's role in sepsis progression, resolution, and long-term outcome', *Immunological Reviews*, 274(1), pp. 330–353. Available at: https://doi.org/10.1111/imr.12499.

[13] Rhodes, A. *et al.* (2017) 'Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016', *Intensive Care Medicine*, 43(3), pp. 304–377. Available at: https://doi.org/10.1007/s00134-017-4683-6.

[14] Liang, H. *et al.* (2021) 'Corticosteroids for Treating Sepsis in Adult Patients: A Systematic Review and Meta-Analysis', *Frontiers in Immunology*, 12, p. 709155. Available at: https://doi.org/10.3389/fimmu.2021.709155.

[15] Annane, D. *et al.* (2016) 'Design and conduct of the activated protein C and corticosteroids for human septic shock (APROCCHSS) trial', *Annals of Intensive Care*, 6(1), p. 43. Available at: https://doi.org/10.1186/s13613-016-0147-3.

[16] Annane, D. *et al.* (2018) 'Hydrocortisone plus Fludrocortisone for Adults with Septic Shock', *New England Journal of Medicine*, 378(9), pp. 809–818. Available at: https://doi.org/10.1056/NEJMoa1705716.

[17] Fleuriet, J. *et al.* (2023) 'Rapid rEcognition of COrticosteRoiD resistant or sensitive Sepsis (RECORDS): study protocol for a multicentre, placebo-controlled, biomarker-guided, adaptive Bayesian design basket trial', *BMJ open*, 13(3), p. e066496. Available at: https://doi.org/10.1136/bmjopen-2022-066496.

[18] Rautanen, A. *et al.* (2015) 'Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study', *The Lancet. Respiratory Medicine*, 3(1), pp. 53–60. Available at: https://doi.org/10.1016/S2213-2600(14)70290-5.

[19] Hernandez-Beeftink, T. *et al.* (2022) 'A genome-wide association study of survival in patients with sepsis', *Critical Care (London, England)*, 26(1), p. 341. Available at: https://doi.org/10.1186/s13054-022-04208-5.

[20] Niemi, M.E.K. *et al.* (2021) 'Mapping the human genetic architecture of COVID-19', *Nature*, 600(7889), pp. 472–477. Available at: https://doi.org/10.1038/s41586-021-03767-x.

[21] Chapman, S.J. and Hill, A.V.S. (2012) 'Human genetic susceptibility to infectious disease', *Nature Reviews. Genetics*, 13(3), pp. 175–188. Available at: https://doi.org/10.1038/nrg3114.

[22] Scherag, A. *et al.* (2016) 'Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality', *EBioMedicine*, 12, pp. 239–246. Available at: https://doi.org/10.1016/j.ebiom.2016.08.043.

[23] Rosier, F. *et al.* (2021) 'Genetic Predisposition to the Mortality in Septic Shock Patients: From GWAS to the Identification of a Regulatory Variant Modulating the Activity of a CISH Enhancer', *International Journal of Molecular Sciences*, 22(11), p. 5852. Available at: https://doi.org/10.3390/ijms22115852.

[24] D'Urso, S. *et al.* (2020) 'Septic Shock: A Genomewide Association Study and Polygenic Risk Score Analysis', *Twin Research and Human Genetics: The Oficial Journal of the International Society for Twin Studies*, 23(4), pp. 204–213. Available at: https://doi.org/10.1017/thg.2020.60.

[25] Guillen-Guio, B. *et al.* (2020) 'Sepsis-associated acute respiratory distress syndrome in individuals of European ancestry: a genome-wide association study', *The Lancet. Respiratory Medicine*, 8(3), pp. 258–266. Available at: https://doi.org/10.1016/S2213-2600(19)30368-6.

[26] Kwok, A.J., Mentzer, A. and Knight, J.C. (2021) 'Host genetics and infectious disease: new tools, insights and translational opportunities', *Nature Reviews. Genetics*, 22(3), pp. 137–153. Available at: https://doi.org/10.1038/s41576-020-00297-6.

[27] Kosmicki, J.A. *et al.* (2021) 'Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals', *American Journal of Human Genetics*, 108(7), pp. 1350–1355. Available at: https://doi.org/10.1016/j.ajhg.2021.05.017.

[28] Povysil, G. et al. (2021) 'Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19', The Journal of Clinical Investigation, 131(14), p. e147834. Available at: https://doi.org/10.1172/JCI147834.

[29] Zhang, Q. et al. (2020) 'Inborn errors of type I IFN immunity in patients with life-threatening COVID-19', Science (New York, N.Y.), 370(6515), p. eabd4570. Available at: https://doi.org/10.1126/science.abd4570.

[30] Kousathanas, A. et al. (2022) 'Whole-genome sequencing reveals host factors underlying critical COVID-19', Nature, 607(7917), pp. 97–103. Available at: https://doi.org/10.1038/s41586-022-04576-6.

[31] Zeberg, H. and Pääbo, S. (2020) 'The major genetic risk factor for severe COVID-19 is inherited from Neanderthals', Nature, 587(7835), pp. 610–612. Available at: https://doi.org/10.1038/s41586-020-2818-3.

[32] Berlin, D.A., Gulick, R.M. and Martinez, F.J. (2020) 'Severe Covid-19', New England Journal of Medicine, 383(25), pp. 2451–2460. Available at: https://doi.org/10.1056/NEJMcp2009575.

[33] Schork, N.J. *et al.* (2009) 'Common vs. Rare Allele Hypotheses for Complex Diseases', *Current opinion in genetics & development*, 19(3), pp. 212–219. Available at: https://doi.org/10.1016/j.gde.2009.04.010.

[34] Fallerini, C. *et al.* (2022) 'Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity', *Human Genetics*, 141(1), pp. 147–173. Available at: https://doi.org/10.1007/s00439-021-02397-7.

[35] Danecek, P. et al. (2021) 'Twelve years of SAMtools and BCFtools', GigaScience, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.