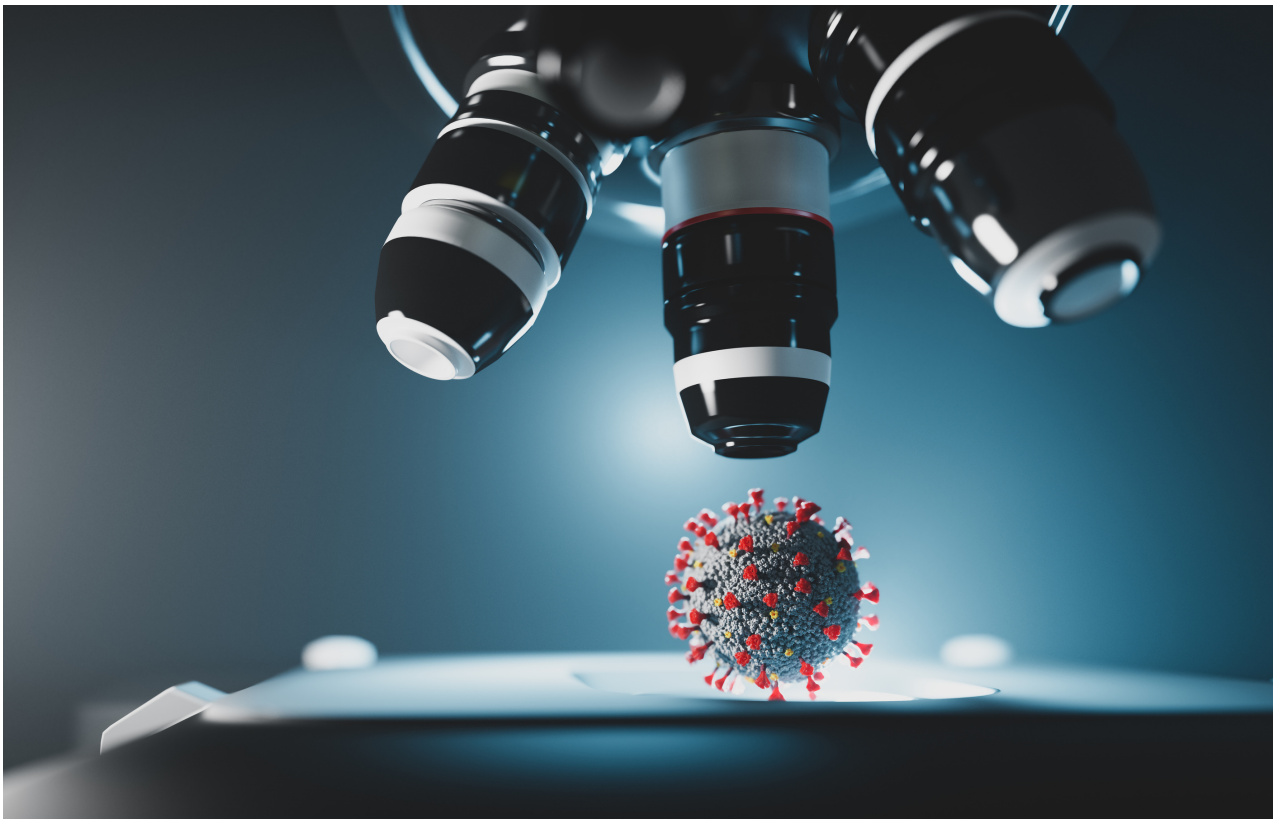
Interreg
Greece-Italy
CoofHea
European Regional Development Fund

EUROPEAN UNION



Study of genetic variants related to the **SARS-CoV-2 host response (INTERACTOME)** ***in Apulia Region***

Authors: Nicoletta Resta and Graziano Pesole
University of Bari "A. Moro"

Study of genetic variants related to the SARS-CoV-2 host response (INTERACTOME) in Apulia Region

Authors: Nicoletta Resta and Graziano Pesole
University of Bari "A. Moro"

Year of Release: 2022

Scientific Report produced within the framework of the **Project CoofHea – Cooperation for Health**


Foggia
Barletta
Andria
Trani
Bari
Taranto
Brindisi
Lecce
ITALY
GREECE
Ioannina
Thesprotia
Arta
Preveza
Kerkyra
Aitoloakarnania
Lefkada
Achaia
Kefallinia
Ileia
Zakynthos

TABLE OF CONTENT

THE PROJECT

The Covid-19 pandemic is undoubtedly an obvious threat to public health that requires a huge commitment by the Italian, Greek and international scientific community aimed at a deeper knowledge of both the molecular mechanisms that cause the pathological condition and the evolutionary characteristics of the virus genome pathogen SARS-CoV-2.

The COVID-19 outbreak has affected Member States in a sudden and dramatic manner and will have implications, specifically on the Greece Italy Programme territory. In Puglia, at the end of July 2020, there were over than 4,500 cases of COVID-19 (1,14‰ of the total population). According to the official data, 551 people have died for COVID-19 causes in Puglia. The profile of the sick people has an average age of 56 years out of which 51,2% are of male gender. In Greece, in the same period, there are 4,401 cases 20 died people for COVID-19 causes.

With these data, it is clear the involvement of all the eligible area. Until today we are living a great impact of the COVID-19 and it is necessary to have a common approach in order to try to solve or to mitigate the problem.

The COOFHEA project, indeed, will operate with three approaches:

A. supporting Puglia Region and the Hospitals in the Greek eligible area to purchase the personal protective equipment and/or medical equipment (ventilators, beds, monitors, etc.).
B. supporting scientific research that will use a set of methods based on different types of approaches, to identify possible molecular mechanisms that can be exploited for the development of innovative and more efficient therapies.
C. sharing with the Greek hospitals the Puglia platform of telemedicine/clinical remote assistance. This is an innovative system with lets the monitoring of patients forced to quarantine for Covid-19 in their home, in order to avoid the hospitals costs, thanks to the "H-Casa" clinical remote assistance platform and to give a possibility of "normal life" to the population affected but not in a very serious way.

With this differentiated approach it will be possible to give organic answers to a complex problem like COVID-19.

BRIEF
INTRODUCTION

Starting from March 2020, many research groups worldwide have been engaged in the identification and functional characterization of human genetic variations associated with different levels of susceptibility to SARS-CoV-2 infection and different clinical manifestations of COVID19, see Velavan et al, 2021 and Severe Covid-19 GWAS Group et al. (2020) for an up to date review.

The clinical course of SARS-CoV-2 infection varies greatly amongst individuals, ranging from asymptomatic, to a mild illness and up to a deadly disease (Hu et al, 2021). Since 1950, genetic and molecular research has established an immunological basis for inherited susceptibility to infectious illnesses. Autosomal recessive neutropenia and X-linked recessive agammaglobulinemia were discovered through patient and family research. The pathophysiological mechanism underlying each of these major inborn defects of immunity was identified, providing proof of principle for a genetic basis to the different levels of susceptibility of human beings to infectious disorders. Since 1985, molecular genetics research has revealed that the majority of pathological conditions associated with an impaired immune response are Mendelian, monogenic and with complete clinical penetrance. However, in the case of COVID-19, several authors report a more complex pattern of modulation of the severity of the disease, which is determined by the interplay between, genomic, epigenomic and environmental factors.

There is no doubt however, that being each individual a unique genotypic framework, genetic determinants are likely to play an important part to the likelihood of becoming seriously or mildly ill. The majority of SARS-CoV-2 infections are asymptomatic or benign, but SARS-CoV-2 infectious disease 2019 (COVID-19) can cause life-threatening disease, which usually begins with pneumonia. In the critical sickness induced by COVID- 2019, host-mediated pulmonary inflammation is present and may drive to death. Severe COVID-19 is more common in people over the age of 50, as well as in those who have comorbidities such as pulmonary, cardiovascular, and metabolic disorders.

The relatively recent advent of modern sequencing technologies represented a major breakthrough in contemporary human genetics. By allowing an accurate and complete investigation of the genetic make-up of a single individual, these technologies have helped researchers in building/collecting a comprehensive catalogue of human genetic variation, and most important in the identification of genetic traits or alleles associated with different phenotypic conditions. Since each

individual carries in the excess of 4 Million of small indels and/or single nucleotide

variants, the analysis of human genome sequencing data is a highly complex task, and the identification of those variants associated with the disease/phenotype of interest is often compared to the process of finding a needle in a haystack.

In the light of these considerations, experiments for the identification of genetic variants potentially associated with a condition of interest need careful and specific experimental design. For example, in the case of COVID-19, studies of individuals remaining uninfected despite viral exposure and healthy young patients with life-threatening disease without comorbidities represents a possible opportunity to disclose human genetic determinants of infection and disease.

For example, by applying this approach, recent studies by the HGI -Host Genetics Initiative have identified several regions of the genome associated with susceptibility to SARS-CoV2 infection or to increased disease severity. In particular, these regions include:

A. chromosome 12q24.13 in a gene cluster that encodes antiviral restriction enzyme activators (OAS1, OAS2 and OAS3) with an effect on risk of severity of 14-26% in presence of one copy of risk-increasing allele;
B. chromosome 19p13.2 proximal to the gene that encodes tyrosine kinase 2 (TYK2);
C. chromosome 19p13.3 within the gene that encodes dipeptidyl peptidase 9 (DPP9);
D. and chromosome 21q22.1 in the interferon receptor gene IFNAR2.

Other potentially actionable susceptibility loci have been identified on chromosome 9 and 3 (ABO blood group loci and SLC6A20 gene).

In this context the main objectives of our study are to complement/extend previous findings and to identify novel loci and mutations potentially associated with different levels of severity of the clinical manifestation of COVID-19. To this end we have performed exome sequencing of more than 230 subjects living in the Apulia region, each assigned to one of 3 main groups:

1. controls, that is individuals with a mild disease severity;
2. susceptible, individuals with an increased severity of the disease,
3. resistors, individuals resistant to infection or with very mild symptoms. To avoid ascertainment bias and possible confounding effects alle the groups were

matched by age and numerosity.

At the time of sampling the SARS-Cov-2 alpha variant (B.1.1.7) was largely predominant in the Apulia region. For this reason, we did not carry out SARS-Cov-2 genome sequencing in the investigated infected subjects, although viral sequencing was among the initial objectives of the project.

By comparing the profiles of genetic variation across and between groups we identify several point mutations, each specific of a distinct group and which seem to be associated with diverse clinical manifestations of COVID-19 in our cohort. Although several of these genetic variants have not been reported by previous studies, we observe a general enrichment of “candidate mutations” in genes/groups of genes that were already implicated in immune response related phenotypes.

Interestingly, we also observe more general/high level patterns of variation that potentially hint at a widespread epigenetic/transcriptional regulation of human genes implicated in the response to SARS-CoV-2 infection. While additional analyses will be required to extend and eventually validate our findings, the results outlined in this technical report highlight some potential important human genetic variants associated with the modulation of COVID-19 severity.

Ethics declaration: The study was performed in accordance with the Declaration of Helsinki.
Written informed consent was obtained from patients and family members to perform exome analysis for research purposes on peripheral blood or other samples, according to the local ethic committee’s policy (approval code study: **6352 N°0030641/22/04/2020**, Policlinico of Bari, Italy).

EXPERIMENTAL DESIGN
AND SAMPLE POPULATION

Genetic profiles of the individuals enrolled in the study were determined by exome sequencing, by using the Illumina DNA TruSeq Exome Kit . Library preparation and construction was performed according to manufacturer's instructions. Sequencing was performed in 3 distinct runs on a Illumina Novaseq 6000 platform. Exome sequencing, that is targeted re-sequencing of all of the human exons, is currently considered the method of choice for clinical genetics. By targeting only specific portions of the genome that are associated with a known function, this technique can provide a high level of sequencing depth at a relatively reduced cost for the target regions and at the same time facilitate downstream analyses for the functional annotation/interpretation of patterns of genetic variation. In the light of these considerations, exome sequencing was considered the most appropriate strategy for this study.

An average of more than 22 Million reads was obtained for every sample (Table S1). Theoretically, this corresponds to a more than 120x coverage of the target regions. Quality assessment and pre-processing of the data was performed by applying the guidelines defined by our group in Chiara and Pavesi, 2017. Variant calling, for the identification of genetic variants was performed by applying the GATK workflow (Van der Auwera and Connor, 2020) according to the current best practices (as of August 2021). A schematic representation is provided in Figure 1. All the analyses were performed on the hg19 assembly of the human genome. Repeats, alternative haplotypes and pseudo autosomal regions of sexual chromosomes were masked, according to the GATK best practices recommendations. Mapping was performed by applying the bwa-mem software (Li, 2013). Only read pairs with a mapping quality of Q20 or above were considered in all downstream analyses. Post mapping quality metrics were collected by using custom Perl scripts. More than 80% of the reads mapped for every sample covered the panel target regions (Figure 2, panel A). The observed level of coverage was 70x or higher for all the samples (Figure 2, panel B). In line with these expectations all the samples achieved a level of coverage of 20x or more for more than 95% of the target regions (Figure 2, panel C) and hence meet the minimal requirements for clinical genetics applications (Bao et al, 2014). Regions that were not covered by 20 or more independent reads, in more than 20 distinct samples were excluded from downstream analyses. Collectively these regions account for 2.66% of the regions targeted by the DNA TruSeq Exome and interestingly they correspond almost perfectly with regions of “low mappability”, as defined by Lee and Schatz, 2012, (data not shown).

Analysis of genetic profiles of the patients and identification of genetic variants potentially implicated in the modulation of COVID-19 severity was performed by using utilities and methods implemented by VINYL (Chiara et al, 2020). Functional annotation was performed by the Annovar software (Wang et al, 2010), which is available in VINYL along with an extensive collection of resources for the annotation of genetic variants (see Table S2). Burden analysis for the identification of genes associated with an excess of regulatory mutations was performed by means of a dedicated tool implemented by VINYL.

A comprehensive list of genes potentially associated with diverse clinical manifestation of COVID-19 was compiled by complementing the expert curated panel of genes potentially implicated in the modulation immune-disease severity provided by Genomics England (as available from https://panelapp.genomicsengland.co.uk/panels/111/) with further manual curation of recently published studies.

The sample population subjected to exome sequencing included 236 individuals recruited respectively in intensive and sub-intensive care units (Policlinico di Bari) as well as in the Apulian territory directly to the patients' homes through the USCA services which provided direct assistance to COVID-19 patients. Studies in literature (Buzdugan et al, 2016), suggest that this sample size is required to ensure adequate statistical power in carrying out association tests.

For each individual, we collected blood and rhino-oropharyngeal samples, signed informed consent as well as detailed information's regarding age, sex, possible presence of comorbidities such as obesity, hypertension, chronic kidney disease, diabetes, immune compromised status, smoking, cardiovascular disease and chronic respiratory disease. A brief outline of the metadata collected by our study is presented in Figure 3 (panel A).
Each individual was assigned to one of the following groups:

- **susceptible**: young patients infected with severe symptoms and no comorbidities;
- **resistors**: individuals who were exposed to the infection while living in the same household as an infected individual and, preferably, the spouse of an affected individual but did not become infected
- **controls:** individuals not hospitalized, with no symptoms or minor symptoms

As outlined in Figure 3 panel B all the groups were matched in composition, age range and size.

RESULTS AND **DISCUSSION**

Different tests and approaches were applied to identify/characterize genetic variants potentially implicated in the susceptibility to COVID-19. Conceptually our analyses can be broadly categorized into 3 main types:

1. Identification of highly "damaging" variants at disease related genes, by using simple filters and manual curation
2. Identification of highly "damaging" mutations at genes not previously associated with COVID-19, by variant filtration
3. Identification patterns of variation specific to resistors and susceptible individuals by gene mutation burden analyses

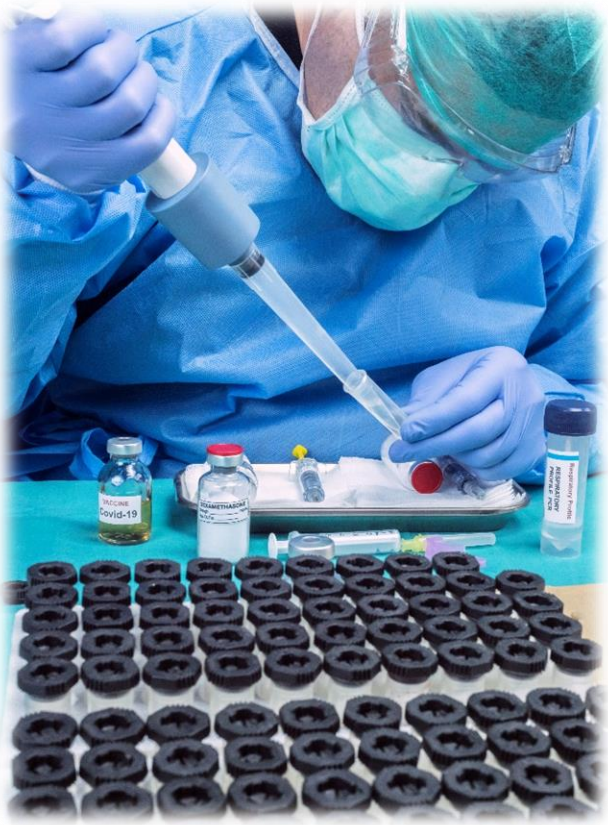
1. Identification of highly "damaging" variants at disease related genes

A comprehensive catalogue of genes implicated in the pathophysiology of COVID-19 was compiled by integrating a expert curated panel provided by Genomics England (as available at: https://panelapp.genomicsengland.co.uk/panels/111/) with further manual curation of the literature. A total of 461 genes potentially associated with the severity of COVID-19 were identified (Table S3)

As outlined in Figure 4 these 461 genes, when principal component analysis of the genetic profiles of the individuals included in this study was carried out, were clearly separated, and susceptible individuals formed a clearly distinct group (first principal component). The second principal component (Y axis of Figure 4), delineated a clear separation between male and female subjects.
Taken all-together this exploratory analysis suggests a distinct pattern of genetic variation between susceptible individuals, and all the other groups included in our study.

Genetic variants were annotated with Annovar (Wang et al, 210). Simple filters were applied to retain only disruptive mutations (i.e frameshift insertions/deletions; splice donor/acceptor mutations and premature stop mutations). Profiles of disruptive mutations were compared and a list of genes specifically disrupted in the S (susceptible) and R (resistor) groups were obtained (Table 1). A total of 4 and 2 genes with disruptive mutations were identified in the S and R group respectively. Interestingly, among the genes specific to the S group, the IFIH1 gene which encodes the MAD5 protein was previously reported to be an important effector of the detection of SARS-CoV-2 infection (Sampaio et al, 2021), while the ZNF341 is the master regulator of STAT3, a member of the STAT family of genes that mediates

cellular responses to interleukins interferon and other growth factors. Interestingly,



defects in response to interferon signaling have already been implicated in severe clinical manifestations of COVID-19 (see Lee and Shin 2020 for a review). The other 2 genes associated with disruptive mutations in the S group, IL17RC and CFTR have already been associated with reduced immune response and a higher propensity to chronic infection (Lyczak et al, 2002) however to the best of our knowledge a direct link with COVID-19 has not been established yet.

Conversely, RNASEL and ORAI1, the two genes in our short-list associated with disruptive mutations in the R (resistor) group, do not seem to be strongly associated with defects in the immune system and further investigations will be required to understand their potential impact on the etiopathology of COVID-19.

2. Identification of highly "damaging" variants exome-wide

The same approach described in the previous section was applied to identify genes associated with highly disruptive mutations only in the R or S groups, without any filter for the selection of disease related genes. A total of 3,298 genes were found to have at least one or more disruptive mutation in at least one of our 236 subjects. Of these 38 were associated with one or more disruptive mutations only in S; the equivalent number for the R group is 26.

Consistent with our previous observations, while R group genes did not seem to be associated with immune-system in general, a slight over-representation of immune system related genes was observed in the S group. For example, apart from ZNF341, IFIH1, IL17RC and CFTR the group S specific list includes genes that implicated in the regulation of the immune response, such as GSN, and GBP2 or genes associated with

phagocytosis (ATP13A2, WIPI2).

In the light of these observations, our data suggest that while disruption of immune system related genes might be a predisposing condition for the development of more severe form of COVID-19, no specific functional enrichment is observed for genes carrying deleterious mutations in subjects resistant to the infection of SARS-CoV-2.

3. Burden analysis and identification of genes showing an excess of mutations

We speculated that different molecular mechanisms, despite the disruption of protein coding genes, could be associated with the modulation of the response to SARS-CoV-2 infection.

For example, transcriptional regulation and other epigenetic mechanisms could impair/or boost immune response by modulating gene expression patterns. Mutational burden analysis of known regulatory variants is a common approach to evaluate (indirectly) the extent/effect of regulatory and epigenetic mechanisms in genotyping studies (Gibson et al, 2015). By this method, genes associated with an increased or reduced expression in any group of interest, can be identified by comparing mutation profiles of the cohort under study with databases and resources of genetic variants and/or haplotypes associated with regulatory effects. Where available, tissues and/or cellular types of interest can be selected to enhance the specificity of the results.

The GTEx portal was used to retrieve a collection of expression Quantitative Trait Loci (eQTLs), that is genomic

loci that explain variation in expression levels of mRNAs in pulmonary tissues (GTEx Consortium, 2013). Burden analyses were performed to identify genes showing an excess of variants associated with eQTLs in the R and S groups, compared with the controls.

A total of 37 and 27 genes respectively were found to be associated with an excess of eQTLs in S and R respectively. Functional enrichment analysis by the EnrichR tool, suggested a highly significant enrichment of gene ontology terms associated with immune system development and KEGG pathways related to Interferon signalling (Figure 5 A and B) in genes specific to the S group. While genes associated with an excess of eQTLs in the R group (Figure 6A and 6B) recovered a somewhat different pattern of enrichment, which consistent with our previous observations does not seem to be related/associated with immune system genes.

Interestingly, however, we notice (Figure 5C and Figure 6C), that both sets are very highly enriched in genes differentially expressed upon SARS-CoV-2 infection in different model systems. An observation that provides an indirect validation of the approach applied in our study and which potentially warrants further investigations.

CONCLUSIONS

By performing extensive analyses of genetic profiles of individuals affected by COVID-19, this study identified some novel, potentially interesting candidate genes associated with different clinical manifestations of the disease (see Table 1 and Table 2). In line with previous studies, we observe that the majority of the allelic variants associated with an increased susceptibility to COVID-19 seem to be associated with immune-system genes, or genes involved in the modulation of the immune-response.

Importantly, apart from extending the catalog of currently known variants and genes associated with more severe forms of COVID-19, our analyses also suggest that genetic variants associated with an increased susceptibility to the disease are not limited highly disruptive variants in protein coding genes, but might also include regulatory and epigenetic variants which concur in the modulation of the complex pattern of expression of immune genes (Table 3).

While the interpretation of variants associated with increased susceptibility to COVID-19 in our study was relatively simple, although a similar number of genes and variants potentially associated with resistance to the disease was identified, the functional interpretation of these results does not seem to be as straightforward. Indeed, although the majority of the genes associated with mutations in the R (resistor) group in our study, is differentially expressed upon SARS-CoV-2 infection, and in different model systems (Figure 6C), functional annotation of these genes and manual queries of the literature did not provide a reasonable direct link to possible mechanistic insights. While these results might suggest/be related to:

1. a limited knowledge of the function of many genes in our genome;
2. a limited applicability of the standard methods for the prioritization of genetic variants for the identification of markers associated with resistance to SARS-CoV-2 infection; or (more likely)
3. the involvement of environmental and epigenetics factors.

In conclusion we believe that our data reveal the existence of a variability in the individual's response to infection, which could reside in a different combination of genetic, epigenetic and environmental factors. In the light of the above considerations, we foresee that the integration of different type of omics and datasets will be required in the future to identify the complex layers of variation that subtend different clinical manifestations of COVID-19.

REFERENCES

~ Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. A., Jiang, H., & Feng, G. (2014). Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. Cancer informatics, 13(Suppl 2), 67–82. https://doi.org/10.4137/CIN.S13779

~ Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., & Bühlmann, P. (2016). Assessing statistical significance in multivariable genome wide association analysis. Bioinformatics (Oxford, England), 32(13), 1990–2000. https://doi.org/10.1093/bioinformatics/btw128

~ Chiara, M., & Pavesi, G. (2017). Evaluation of Quality Assessment Protocols for High Throughput Genome Resequencing Data. *Frontiers in genetics*, *8*, 94. https://doi.org/10.3389/fgene.2017.00094

~ Chiara, M., Mandreoli, P., Tangaro, M. A., D'Erchia, A. M., Sorrentino, S., Forleo, C., Horner, D. S., Zambelli, F., & Pesole, G. (2020). VINYL: Variant prIoritizatioN by survivaL analysis. *Bioinformatics (Oxford, England)*, btaa1067. Advance online publication. https://doi.org/10.1093/bioinformatics/btaa1067

~ GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nature genetics, 45(6), 580–585. https://doi.org/10.1038/ng.2653

~ Gibson, G., Powell, J. E., & Marigorta, U. M. (2015). Expression quantitative trait locus analysis for translational medicine. Genome medicine, 7(1), 60. https://doi.org/10.1186/s13073-015-0186-7

~ Hu, B., Guo, H., Zhou, P., & Shi, Z. L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nature reviews. Microbiology, 19(3), 141–154. https://doi.org/10.1038/s41579-020-00459-7

~ Lee, H., & Schatz, M. C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics (Oxford, England), 28(16), 2097–2105. https://doi.org/10.1093/bioinformatics/bts330

~ Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2

~ Lyczak, J. B., Cannon, C. L., & Pier, G. B. (2002). Lung infections associated with cystic fibrosis.
Clinical microbiology reviews, 15(2), 194–222.
https://doi.org/10.1128/CMR.15.2.194-222.2002

~ Pairo-Castineira, *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* 591, 92–98 (2021).

~ Sampaio, N. G., Chauveau, L., Hertzog, J., Bridgeman, A., Fowler, G., Moonen, J. P., Dupont, M., Russell, R. A., Noerenberg, M., & Rehwinkel, J. (2021). The RNA sensor MDA5 detects SARS-CoV-2 infection. Scientific reports, 11(1), 13638.
https://doi.org/10.1038/s41598-021-92940-3

~ Severe Covid-19 GWAS Group et al. (2020) Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med*. 2020;383(16):1522-1534.

~ Thakur B. A systematic review and meta-analysis of geographic differences in comorbidities and associated severity and mortality among individuals with COVID-19. Sci Rep. 2021;11(1):8562.
Published 2021 Apr 20. doi:10.1038/s41598-021-88130-w

~ Velavan, T. P., Pallerla, S. R., Rüter, J., Augustin, Y., Kremsner, P. G., Krishna, S., & Meyer, C. G. (2021). Host genetic factors determining COVID-19 susceptibility and severity. *EBioMedicine*, *72*, 103629.
https://doi.org/10.1016/j.ebiom.2021.103629

~ Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. 2010 Nucleic Acids Research, 38:e164.

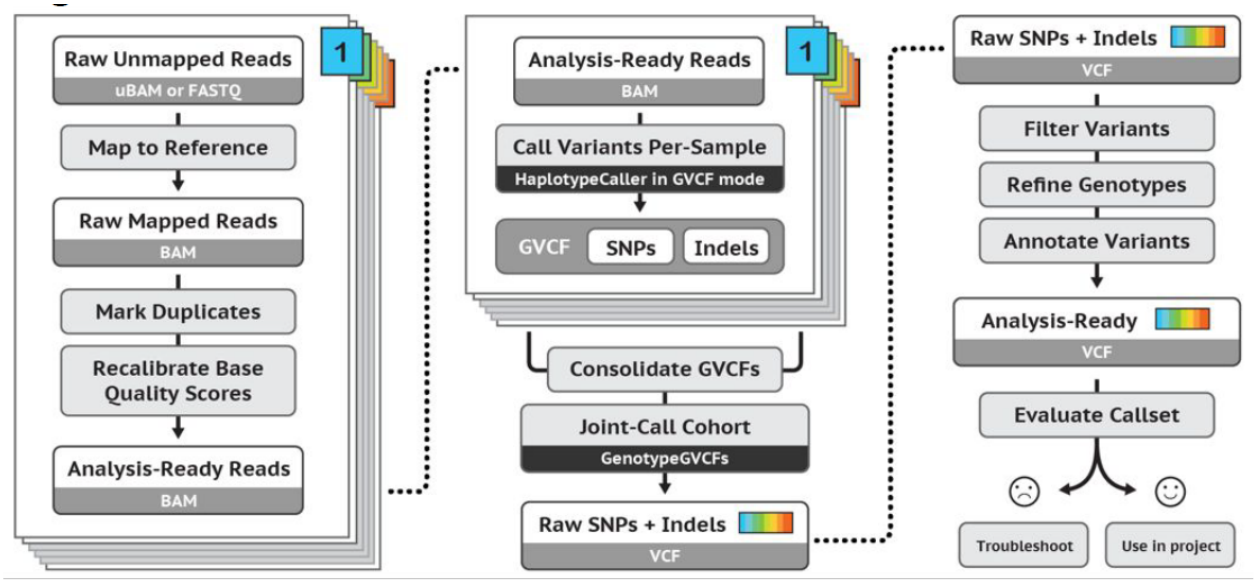
~ Van der Auwera GA & O'Connor BD. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media.

~ Zhang Q, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19.
Science. 2020;370(6515):eabd4570. doi:10.1126/science.abd4570

FIGURES **LEGENDS**

Figure1: Schematic of the GATK workflow

Adapted from:
https://gatk.broadinstitute.org/hc/en-us/articles/360035890411-Calling-variants-on-cohorts-of-samples-using-the-HaplotypeCaller-in-GVCF-mode .



Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

This is the recommended workflow for performing variant discovery analysis on cohorts of samples. Briefly 1) Reads are pre-processed and aligned to the reference genome according to the GATK best practices. 2) Each sample is genotyped by using the HaplotypeCaller. Genetics profiles in GVCF format are obtained. 3) Genetic profiles of all the individuals are aggregated into a single GVCF. 4) Variant recalibration is applied for the identification of the final call-set.

Figure 2: Boxplot of quality metrics

A) % of regions of target
B) Theoretical coverage
C) % of target regions covered by 20 or more reads. In each plot the dotted read lines represent the ideal/expected value for every metric.

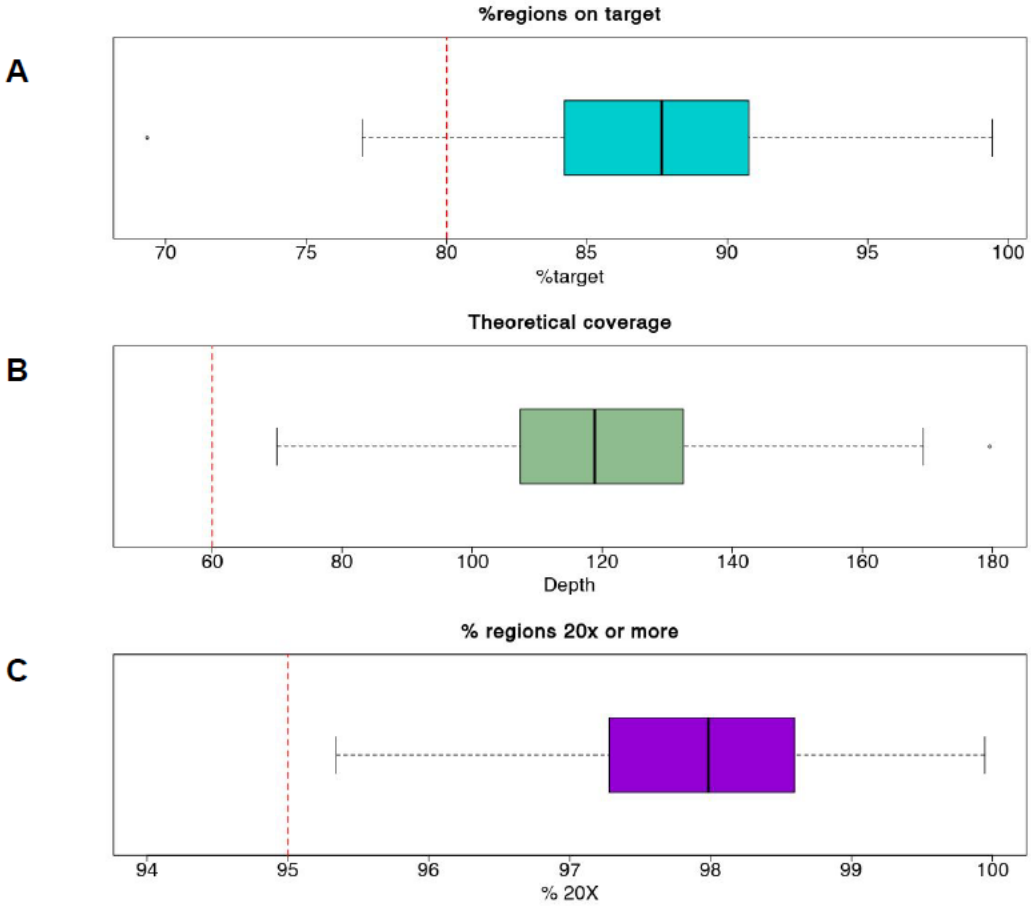


Figure 3: Metadata collection and breakdown of groups

A) Information collected from every patient.

B) Breakdown of the cohort by age, sex and phenotype.

A

Sampling Collection for Covid-HGE e COOFHEA

ID	Sampling Date	Sample Origin	Sex	Age	Consent	Clinical status	Swab
IT_BA_C0001	02/03/2021	ASL BARI	F	26	yes	healthy	yes
IT_BA_C0004	04/03/2021	ASL BARI	F	25	yes	healthy	yes
IT_BA_C0008	04/03/2021	ASL BARI	M	45	yes	healthy	yes

Comorbidities

hypertension	diabetes	cardiovascular disease	chronic respiratory disease	chronic kidney disease	immune compromised status	obesity	smoking
no	no	no	no	no	no	no	no
no	no	no	no	no	no	no	no

B

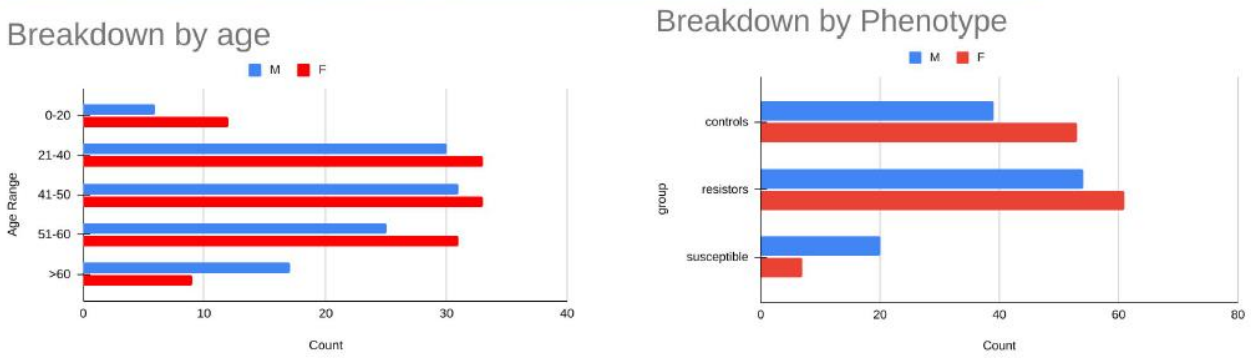


Figure 4: PCA (Principal component analysis) of genetic profiles

Principal component analysis of genetic profiles of the individuals included in our cohort at 462 genes implicated with the modulation of COVID-19 severity. R=resistors (green), S=susceptible (red), C=controls (blue)

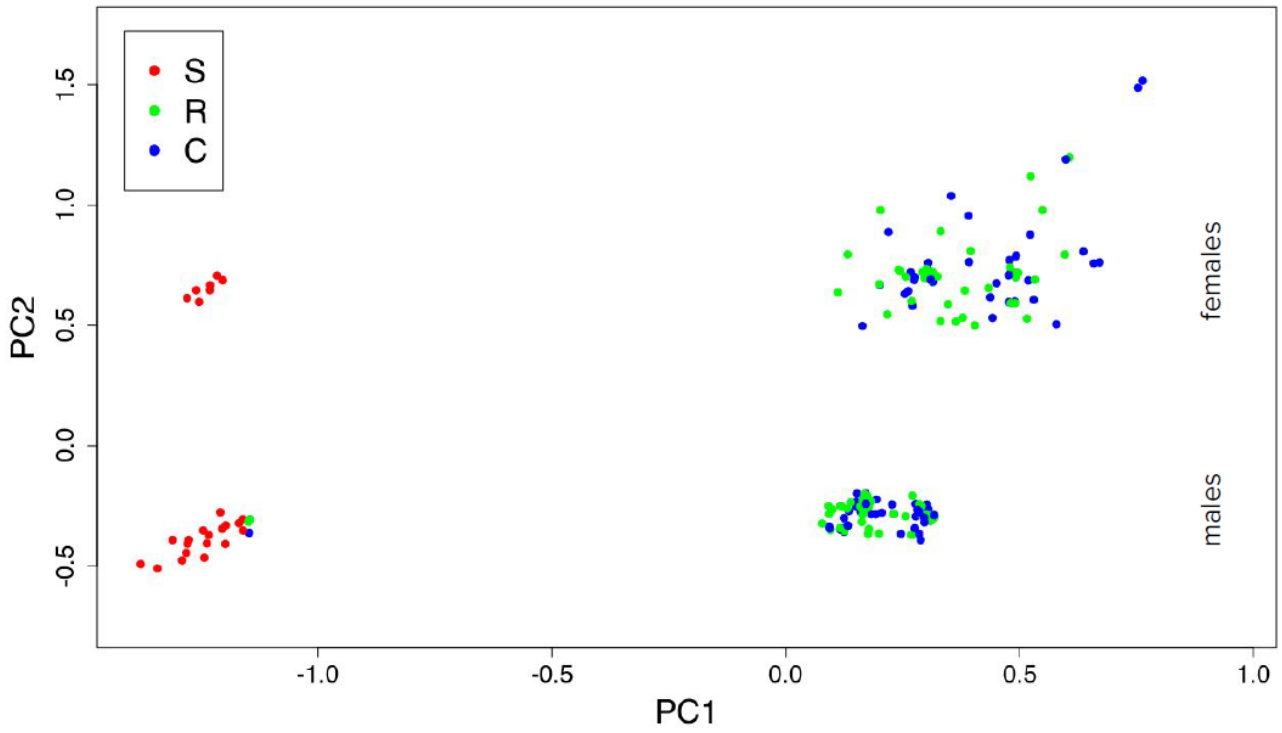


Figure 5: Functional enrichment analysis of genes enriched in eQTLs in the S group

A) Gene ontology (molecular function). B) KEGG pathways. C) genes differentially expressed upon SARS-CoV-2 infection

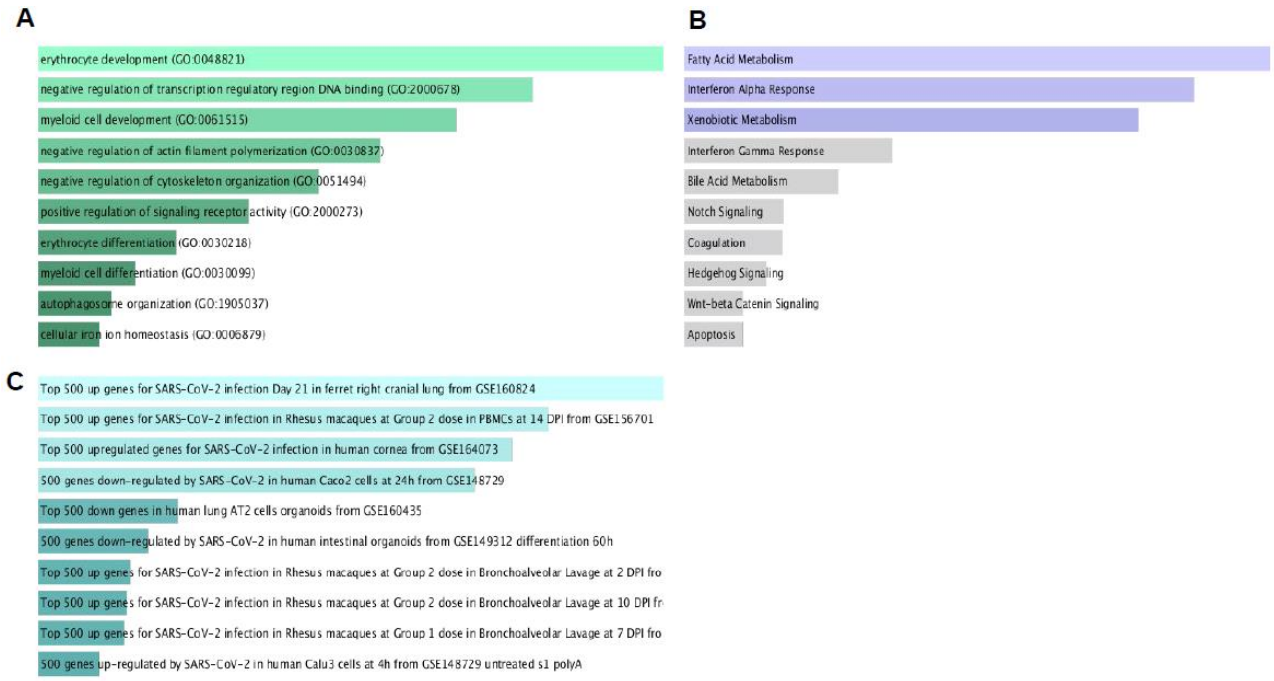


Figure 6: Functional enrichment analysis of genes enriched in eQTLs in the R group

A) Geneontology (molecular function)
B) KEGG pathways
C) genes differentially expressed upon SARS-CoV-2 infection

TABLES **LEGENDS**

Table 1. COVID-19 associated genes with highly disruptive mutations in Susceptible (S) and Resistors (R) individuals.

N° individuals= number of individuals carrying a disruptive mutation in that gene.

Table 1			
Susceptible (S)		**Resistors (R)**	
Gene	**N° of individuals**	**Gene**	**N° of individuals**
ZNF341	2	RNASEL	3
CFTR	2	ORAI1	2
IL17RC	1		
IFIH1	5		

Table 2. Complete list of genes with highly disruptive mutations in Susceptible (S) and Resistors (R) individuals

N° individuals= number of individuals carrying a disruptive mutation in that gene.

Table 2			
Susceptible (S)		**Resistors (R)**	
Gene	**N° of individuals**	**Gene**	**N° of individuals**
FOXD3	2	RNASEL	3
ALAS2	1	ORAI1	2
CDX4	1	C1QTNF3	1
ZNF341	2	UBB	3
FAM89A	1	KCNA3	1
DKC1	1	ADRA2A	2
YRDC	1	NOMO2	2
GBP2	1	MSS51	1
FAM43A	3	PPM1K	4
RTN4RL2	1	OR1L3	1
IQCE	1	PABPN1	1
KANK3	2	GSTT2B	2
TMEM80	3	DCAF4L2	3
GSN	1	HMGN5	2
HTRA1	1	HIST1H4K	1
CTDP1	1	RNF182	1
SHANK3	2	NUDT11	1
SMTNL2	1	BOLA1	2
DRD4	1	C11orf57	1
ZCCHC24	1	OR51T1	2
ACADM	1	SNRNP35	2
WIPI2	1	SMIM1	1

ABCA9	1	MAP6	1
TMEM187	3	CEP78	1
KIF7	1	ASCL1	1
ATP13A2	1	ABHD14B	1
PDZRN3	1	GAS1	1
LGSN	1		
CHM	1		
TSSK6	2		
IRF2BP1	1		
KDM2B	1		
TAF1	3		
KLHL34	1		
CFTR	2		
IL17RC	1		
IFIH1	5		

Table 3. List of genes with a statistically significant over-representation of eQTLs in lung tissues in Susceptible (S) and Resistors (R) individuals

FDR= statistical significance of the over-representation, after the application of the Bonferroni procedure for the control of False Discovery Rate

Table 3			
Susceptible (S)		**Resistors (R)**	
Gene	**FDR**	**Gene**	**FDR**
GP1BB	0,00331301	MRPL40	0,000142487
KDM2B	0,001101893	IRS1	0,000332023
TAF1	0,000578853	CDC42EP5	0,002158521
KLHL34	0,000255155	PRPSAP1	0,001317715
ZCCHC13	0,00016767	ATP6AP1L	0,000492932
GATA1	0,00109483	CPSF3	0,000815716
CDX4	5,68E-05	ALX3	0,000797532
ALAS2	0,001730036	ARX	0,002019728
FOXD3	0,00040674	TUBG1	0,003522145
ZNF341	0,001490209	ZBED3	0,001862392
FAM89A	0,002302069	CIR1	0,000808805
DKC1	0,001627822	CCT8	0,001465839
YRDC	7,61E-05	KNOP1	0,002284595
GBP2	0,002816084	LAMP1	0,000142636
FAM43A	0,00039293	HNRNPA1L2	0,00203004
RTN4RL2	0,00237345	SLC12A4	0,00127649
IQCE	0,001210283	NOX1	0,001224229

KANK3	2,26E-06	CEL	0,000175619
TMEM80	0,000489549	CSNK1G2	0,001497076
GSN	0,000948471	VWCE	2,57E-05
HTRA1	0,003338918	SOAT2	0,001733679
CTDP1	0,000998612	RS1	0,000647716
SHANK3	0,000111655	PTPN18	0,001811143
SMTNL2	0,000714511	CCR2	0,000966215
DRD4	0,001304169	HAS1	0,001005425
ZCCHC24	0,000549543	LYPD6B	0,00102652
ACADM	0,003792577	SLC22A11	0,000500303
WIPI2	0,000707617		
ABCA9	0,003673757		
TMEM187	0,001329245		
KIF7	0,001176848		
ATP13A2	0,001175405		
PDZRN3	0,000122889		
LGSN	0,000039492		
CHM	0,000173916		
TSSK6	0,001769331		
IRF2BP1	0,000304045		

Supplementary Tables Legends

***These tables and all the relevant data can be provided upon request*

Supplementary Table S1: metadata and reads mapping metrics

Metadata collected for every sample are reported in columns A to Q. Reads mapping metrics in columns R to W

Supplementary Table S2: List of resources for the annotation of genetic variants included in VINYL.
First column_ name of the resource according to Annovar.
Second column: brief description of the resource.
Third column: type of Annovar operation associated with the resource.

Supplementary Table S3: List of 461 COVID-19 related genes

Interreg
Greece-Italy
CoofHea
European Regional Development Fund

EUROPEAN UNION

CONTACT INFORMATION

Prof. Nicoletta Resta
Department of Biomedical Sciences and Human Oncology, University of Bari

nicoletta.resta@uniba.it
+39 080 5593619
Piazza G. Cesare 11
70124 Bari, Italy

Prof. Graziano Pesole
Department of Biosciences, Biotechnology and Biopharmaceuticals

graziano.pesole@uniba.it
+39 080 5443588
Via Edoardo Orabona,
70125 Bari, Italy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Project co-funded by European Union, European Regional Development Funds (E.R.D.F.) and by National Funds of Greece and Italy

REGIONE PUGLIA
Πανεπιστημιακό Γενικό Νοσοκομείο Ιωαννίνων
Πανεπιστημιακό Γενικό Νοσοκομείο Πατρών "Παναγία η Βοηθεία"
Γενικό Νοσοκομείο Πατρών "Ο Άγιος Ανδρέας"
Γενικό Νοσοκομείο Ηλείας
ΓΝΑ Γενικό Νοσοκομείο Αιτωλοακαρνανίας
Γενικό Νοσοκομείο Κέρκυρας «Αγία Ειρήνη»
Γενικό Νοσοκομείο Ζακύνθου "Άγιος Διονύσιος"
Γενικό Νοσοκομείο Κεφαλληνίας
6η ΥΠΕ Πελοποννήσου - Ιονίων Νήσων "Άγιος Γεράσιμος"