Deep learning identified genetic variants associated with COVID-19 related mortality

Zihuan Li, Wei Dai, Shiying Wang, Yisha Yao, Heping Zhang*

School of Public Health, Yale University, New Haven, CT, USA

60 College Street, New Haven, CT, 06520-8034, USA

*Corresponding author: Heping Zhang; heping.zhang@yale.edu; +12037855185

Abstract

Analysis of host genetic components provides insights into the susceptibility and response to viral infection such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes coronavirus disease 2019 (COVID-19). To reveal genetic determinants of susceptibility to COVID-19 related mortality, we train a deep learning model to identify groups of genetic variants and their interactions that contribute to the COVID-19 related mortality risk using the UK Biobank data. We refer to such groups of variants as super variants. We identify 15 super variants with various levels of significance as susceptibility loci for COVID-19 mortality. Specifically, we identify a super variant (OR=1.594, p=5.47×10^{-9}) on Chromosome 7 that consists of the minor allele of rs76398985, rs6943608, rs2052130, 7:150989011_CT_C, rs118033050 and rs12540488. We also discover a super variant (OR=1.353, p=2.87×10^{-8}) on Chromosome 5 that contains rs12517344, rs72733036, rs190052994, rs34723029, rs72734818, 5:9305797_GTA_G and rs180899355.

Keywords: Deep learning, COVID-19, SARS-CoV-2, UK Biobank, TAS2R1.

Introduction

Coronavirus disease 2019 (COVID-19) is a pandemic viral disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and has resulted in significant morbidity and mortality worldwide.

The ongoing research efforts to provide new insights into risk of COVID-19 related morbidity and mortality have focused largely on investigating the epidemiological characteristics of COVID-19 [1, 2], and genomic characterization of COVID-19 [3]. According to a cross-sectional survey conducted in the United Kingdom, among the hospitalized COVID-19 patients, those with diabetes, cardiovascular diseases, kidney diseases, obesity, or chronic respiratory diseases are tied to high risk of death [4, 5]. In addition, several recent papers have revealed that males, older people, Black, Asian and Minority Ethnic groups are exposed to an elevated risk of COVID-19 caused mortality [6-

9]. Recent evidence also suggests that host genetic determinants modulate the risk of infection and disease severity [10-13]. Thus, investigating the host genetic basis of heterogeneous susceptibility and severity of COVID-19 and identifying genetic risk factors will deepen our understanding and facilitate the drug developments of COVID-19.

Several genome-wide association studies (GWAS) have been performed to investigate the genetic contribution to COVID-19 outcomes, including the associations between host genetic factors, specifically single-nucleotide polymorphisms (SNPs) and infection or respiratory failure [4, 10, 13, 14], but those studies only focus on the effects of individual SNPs on phenotypes. While most of the literature mentioned above investigate the effects of the variants on COVID-19 susceptibility and severity, few examined association between SNPs and COVID-19 mortality. It is important to examine interactions between SNPs or genes on COVID-19 mortality. Several works utilizing tree-based method have been proposed to overcome drawbacks of traditional GWAS [15-18]. For example, Hu et al. combined a local ranking and aggregation approach with random forest to identify genetic variants associated with risk of COVID-19 related mortality [17]. Though they demonstrated some promising results, including the discovery of some new interacting SNPs, they used an earlier release of the UK Biobank data which contained a much smaller sample size on the COVID-19 cases and deaths.

To account for the potential interactions between SNPs or genes, we adopt the idea of super variants in Song et al [16]. A super variant is a combination of minor alleles in multiple loci throughout a genome in analog to a gene [16]. In contrast to a gene, which is a physically connected region on a chromosome, the loci contributing to a super variant might be located anywhere in the genome. Recent studies have demonstrated the usefulness of super variants in detecting genetic associations with complex diseases [15, 17, 18]. To incorporate potentially complex interactions more effectively among SNPs or genes, we propose a novel approach, namely deep learning-based ranking and aggregation method for identifying genetic variants (DRAG).

To date, deep learning has been successfully applied to numerous tasks in the biomedical domain, such as genomics [19], promoters [20]. Although deep learning methods have been applied in GWAS [21], those methods do not emphasize SNP interaction effects. By contrast, our proposed method is capable of detecting the main and interaction

effects of SNPs that are entangled in high-dimensional and nonlinear representations. In addition, deep learning-based feature selection approaches such as Concrete Auto-encoder (CAE) have been studied by [22, 23]. Abid et al. demonstrated the advantages of CAE: 1) efficacy on a large-scale gene expression dataset; 2) easily scales to high-dimensional datasets; and 3) outperforms a variety of other sophisticated feature selection approaches, including principal feature analysis, multi-cluster feature selection and auto-encoder inspired unsupervised feature selection [23].

The DRAG proceeds in three main steps: SNP-set partition, selecting an optimal subset of SNPs, and determination of super variants. In the first step, SNPs are first divided into consecutive non-overlapping regions with each region consisting multiple SNPs. In the next step, an optimal subset of SNPs within each SNP-set is selected using CAE combined with Bayesian information criterion (BIC) obtained by training logistic regression. Finally, a super variant is formed by the selected subset of SNPs, and the association between the super variant and phenotype is tested by performing logistic regression.

We first apply the DRAG to identify super variants that contribute to the COVID-19 related mortality risk using the UK Biobank white British data to minimize the race and ethnicity impact on our results. We include 18,731 with 8.2 million SNPs as the discovery set. The identified super variants are then tested in an independent validation dataset (n=9,366). Furthermore, we demonstrate the validity and efficacy of our analytic approach through simulation studies.

Results

DRAG identifies super variants associated with COVID-19 related mortality.

We include a total of 28,097 individuals infected with COVID-19 (26,441 survived and 1,656 died) and of white British ancestry from the UK Biobank in our analyses. We consider 8,238,098 SNPs which are divided into 2734 SNP-sets. Each SNP-set is 1Mbp as done in Hu et al. [16, 17] for convenience.

The discovery set contains about two thirds of the participants which lead to the identification of 15 super variants with p-values at or below 1.83×10^{-5} (i.e., 0.05/2734) (Fig. 1).

As the next step, we validate the detected super variants in the remaining one third of the individuals. All of the 15 detected super variants have p-values below 0.05 and one below 0.003 (0.05/15) (Table 1).

Table 1 shows the effects of super variants estimated from logistic regression with sex, age and top 10 principal components (PCs) of the genomic markers [24-26] in the discovery set, as well as the results in the verification and complete sets. The most significant signal for complete set is given by chr4_24 ($p = 2.40 \times 10^{-9}$), and the largest odds ratio appears at chr7_151 (OR = 1.594). The specific formation of the 15 verified super variants and their nearby genes (genes are annotated with ANNOVAR) are given in Table S1 (Supplementary Table 1) where the odds ratio and corresponding *p* values are calculated on the complete dataset.

Supervariants	OR on discovery	p-value on discovery	OR on verification	p-value on verification	OR on complete	p-value on complete
chr2_35	1.401	7.51×10^{-6}	1.289	1.87×10^{-2}	1.368	4.51×10^{-7}
chr4_24	1.464	7.33×10^{-9}	1.214	3.58×10^{-2}	1.376	2.40×10^{-9}
chr4_170	1.381	6.11×10^{-6}	1.303	9.01×10^{-3}	1.354	2.18×10^{-7}
chr5_10	1.343	1.01×10^{-5}	1.383	6.07×10^{-4}	1.353	2.87×10^{-8}
chr5_138	1.361	2.59×10^{-6}	1.207	4.23×10^{-2}	1.306	5.71×10^{-7}
chr6_54	1.395	5.41×10^{-7}	1.212	4.24×10^{-2}	1.329	1.67×10^{-7}
chr6_163	1.439	1.04×10^{-5}	1.342	1.31×10^{-2}	1.406	5.03×10^{-7}
chr7_43	1.398	1.61×10^{-5}	1.329	1.12×10^{-2}	1.370	7.61×10^{-7}
chr7_151	1.688	6.74×10^{-8}	1.419	1.37×10^{-2}	1.594	5.47×10^{-9}
chr10_6	1.359	5.72×10^{-6}	1.217	4.27×10^{-2}	1.309	1.12×10^{-6}
chr11_114	1.364	6.07×10^{-6}	1.242	2.68×10^{-2}	1.321	7.03×10^{-7}
chr13_92	1.374	1.05×10^{-5}	1.226	4.94×10^{-2}	1.320	2.68×10^{-6}
chr16_57	1.425	4.18×10^{-6}	1.341	6.98×10^{-3}	1.394	1.19×10^{-7}
chr17_49	1.379	1.95×10^{-6}	1.209	4.31×10^{-2}	1.314	5.80×10^{-7}
chr20_15	1.333	1.59×10^{-5}	1.267	1.13×10^{-2}	1.310	6.40×10^{-7}

Table 1: Marginal effects of 15 verified super variants on the discovery, verification, and complete dataset using DRAG. The chri_j represents the super variant in the *j*-th set of the *i*-th chromosome; OR on discovery: odd ratio of super variant on discovery set; OR on verification: odd ratio of super variant on verification set; OR on complete: odd ratio of variant on complete set.

Annotation of the identified super variants in their reported roles related to COVID-19

Our analyses identify 4 genetic variants (in/near *TAS2R1*, *ZBTB16*, *LINC01320* and *NCAM1*) [11, 27-31] that have been reported with COVID-19 outcomes.

First, it is worth noting that super variant chr11_114 (OR=1.321, p=7.03×10^{-7}) is composed of 7 SNPs, including the variant rs17115095 in the intronic region of gene *NCAM1* which encodes a cell adhesion protein belonging to the immunoglobulin superfamily. The encoded protein is involved in cell-to-cell and cell-to-matrix interactions during development and differentiation [32]. As a binding partner of spike protein, previous investigations have revealed that there might exist molecular mimicry between *NCAM1* and the COVID-19 envelope protein [29, 30]. Furthermore, in the same super variant, we observe that the SNP rs73000932 is considered as an intronic variant for gene *ZBTB16* (also referred to as promyelocytic leukemia zinc finger). This gene plays a critical role in the function and development of immune system and may enhance T cell responses [33]. Recent studies have reported the upregulation of *ZBTB16* in the tears of COVID-19 patients [31]. These findings suggest that variations within *ZBTB16* is likely to have an effect on the risk of COVID-19 related mortality.

Second, the 7 SNPs in super variant ch5_10 (OR=1.353, p=2.87×10^{-8}), including the SNP index rs180899355, and its nearby gene *TAS2R1* are mapped to chromosome 5p15. This gene encodes a member of the G protein-coupled receptor superfamily that is expressed by taste receptor cells of the tongue and palate epithelia. This intronless taste receptor gene encodes a 7-transmembrane receptor protein, functioning as a bitter taste receptor [34]. Studies have shown that individuals who reported experiencing weak or no bitter tastes were considerably more likely to test positive for COVID-19, to be hospitalized, and to be symptomatic for a longer duration [35]. In addition, recent studies have been conducted that reduced *TAS2R* expression may influence the response to Chloroquine. COVID-19 infected obese patients could respond differently to pharmacological therapy with Chloroquine, with side effects occurring more frequently as result of overdosage [27]. Moreover, the rs180899355 is in the same region as a variant rs148943015 in *TAS2R* gene, specifically rs148943015, which is reported to be associated with COVID-19 susceptibility and severity [11].

In addition, super variant chr2_35 (OR=1.368, p=4.51×10^{-7}) includes SNP rs570706090 whose nearby gene is *LINC01320*. Of note, *LINC01320* expression is reported to be higher in COVID-19 tissue specimens of connecting tubule (CNT) and intercalated cell

type A (IC-A) [28]. In addition, a GWAS study has shown that increased expression of *LINC01320* is associated with testing positive for SARS-CoV-2 [11].

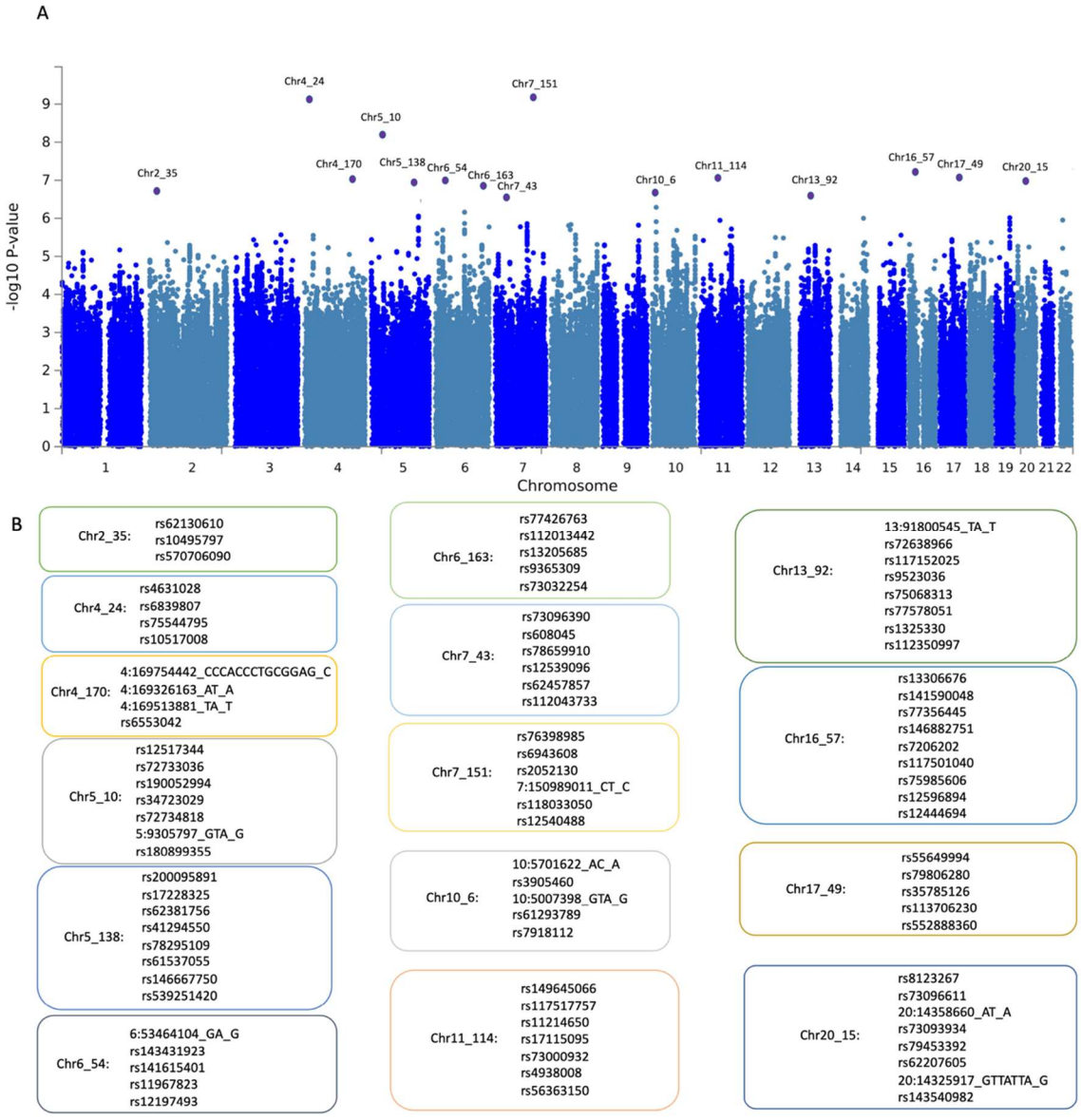


Figure 1: (A) Manhattan plot. **(B)** The set of SNPs in each super variant. The chri_j represents the super variant in the *j*-th set of the *i*-th chromosome.

DRAG identifies novel genes with variants associated with COVID-19 mortality

We also identify 8 novel genes (*DDX60L*, *HSPA9*, *LASTR*, *GLI3*, *AGAP3*, *MACROD2*, *NUP93*, *ELOVL5*) may affect COVID-19 related mortality.

Super variant chr4_170 (OR=1.354, p=2.18×10^{-7}) is composed of 4 SNPs. Variant 4:169326163_AT_A is located in the intronic region of gene *DDX60L* which encodes a protein that is a member of the DEAD-box (DDX) helicase family. Although the function of this gene has not been well characterized, it has been shown that *DDX60L* is involved in anti-viral immunity and acts as a cytosolic sensor of viral nucleic acids [36-38]. The proteins responsible for DDX-mediated hijacking mechanisms are highly conserved among coronaviruses [39]. These observations suggest variations within *DDX60L* and the interaction between them may contribute to COVID-19 disease severity potentially through altered *DDX60L* involving in COVID-19 hijacking mechanisms.

Super variant ch5_138 (OR=1.306, p=5.71×10^{-7}) consists of 8 interacting SNPs. In particular, the variant rs41294550 is within the upstream of gene *HSPA9*. The heat shock protein encoded by this gene is important for cell proliferation, stress response, and mitochondrial maintenance. Recent research indicates that gene *HSPA9* knockdown would result in declined B cells in animal models, while memory B cell levels are associated with protection against COVID-19 delta variant infection [40-42]. These data suggest that variations within *HSPA9* may affect the severity upon COVID-19 infection.

Super variant chr6_54 (OR=1.329, p=1.67×10^{-7}) consists of 4 interacting SNPs, including rs141615401 that lies in the intergenic region of gene *ELOVL5*. The gene *ELOVL5* has been shown to be associated to familial squamous cell lung carcinoma by a previous GWAS study, and several studies indicates that lung diseases would modestly elevate the risk of death after COVID-19 infection [43-45]. In particular, one study notes that patients with lung cancer have more severe symptoms related to COVID-19 infection [52]. Our data suggest that variations within *ELOVL5* and its interaction may result in increased risk of death among COVID-19 patients from lung-related disease.

Furthermore, we capture several SNPs super variants chr7_43 (OR=1.370, p=7.61×10^{-7}), chr7_151 (OR=1.594, p=5.47×10^{-9}) and chr10_6 (OR=1.309, p=1.12×10^{-6}). They are rs7918112 rs78659910 and rs6943608 and their nearby genes are *LASTR*, *GLI3*, and *AGAP3*, respectively. The genes *LASTR* and *AGAP3* have been reported to be associated

with pulmonary function (FEV1/FVC), while the gene *GLI3* has also been found to be related to the pulmonary function (FVC) [46]. COVID-19 virus has been shown to affect many organs, with the lung being one of the most affected [47, 48]. In particular, one study find that severe COVID-19 patients had an abnormal FVC over a half-year following discharge [49]. These findings suggest that future analysis of *LASTR*, *GLI3*, and *AGAP3* gene expression in lung tissues may provide a better understanding on the molecular pathogenesis of COVID-19.

Super variant chr16_57 (OR=1.394, p=1.19×10^{-7}) is composed of 9 SNPs. Among them, rs12596894 is inside an intron of gene *NUP93*. Previous studies reported that severe acute respiratory syndrome coronavirus (SARS-CoV) protein nsp1 disrupts localization of *NUP93* from the nuclear pore complex, and the SARS-CoV-2 virus could have the similar disruption on *NUP93* [50]. These observations suggest that a further investigation on the gene *NUP93* may provide a new insight on COVID-19 pathophysiological mechanisms.

Super variant chr20_15 (OR=1.310 p=6.40×10^{-7}) includes 8 SNPs. They are all located in the intronic regions of nearby gene *MACROD2*. Recent evidence has shown that genetic polymorphisms in the gene *MACROD2* are linked to pulmonary function test parameters [51] , and the *MACROD2* is the closest human homolog of COVID-19 with approximately 30% sequence identity and a similar 3D structure [52]. In addition, the *MACROD2* expression is reported to be associated with the removal of mono (ADP-ribosylation), and COVID-19 encodes a nonstructural protein 3 (nsp3) that contains an ADP-ribose phosphatase domain for immune response [53]. Therefore, our results suggest that variations within *MACROD2* is likely associated with COVID-19 related infection and inflammation, potentially due to alterations in the immune response regulating infection and inflammation.

Comparison with traditional GWAS

We compare super variants identified by DRAG with SNPs identified by traditional GWAS. Again, the same white British ancestry samples with sex, age and 10 PCs are used for GWAS. Through the traditional method, we identify 5 loci (Fig. 2 A) associated with COVID-19 mortality with commonly used threshold of 5×10^{-8}. We find an association at chromosome 5q32, with the index SNP rs7447800 ($p = 2.66 \times 10^{-8}$). This index SNP is in the same region as a functional variant in JAKMIP2 gene, specifically rs10477338, which

is reported to be associated with testing positive for COVID-19 by a GWAS study conducted on approximately1,000,000 individuals [11]. Fig. 2B shows the location of variants on chromosome 5 identified by both methods and suggests that both strategies should be used in tandem to boost the strength of each method.

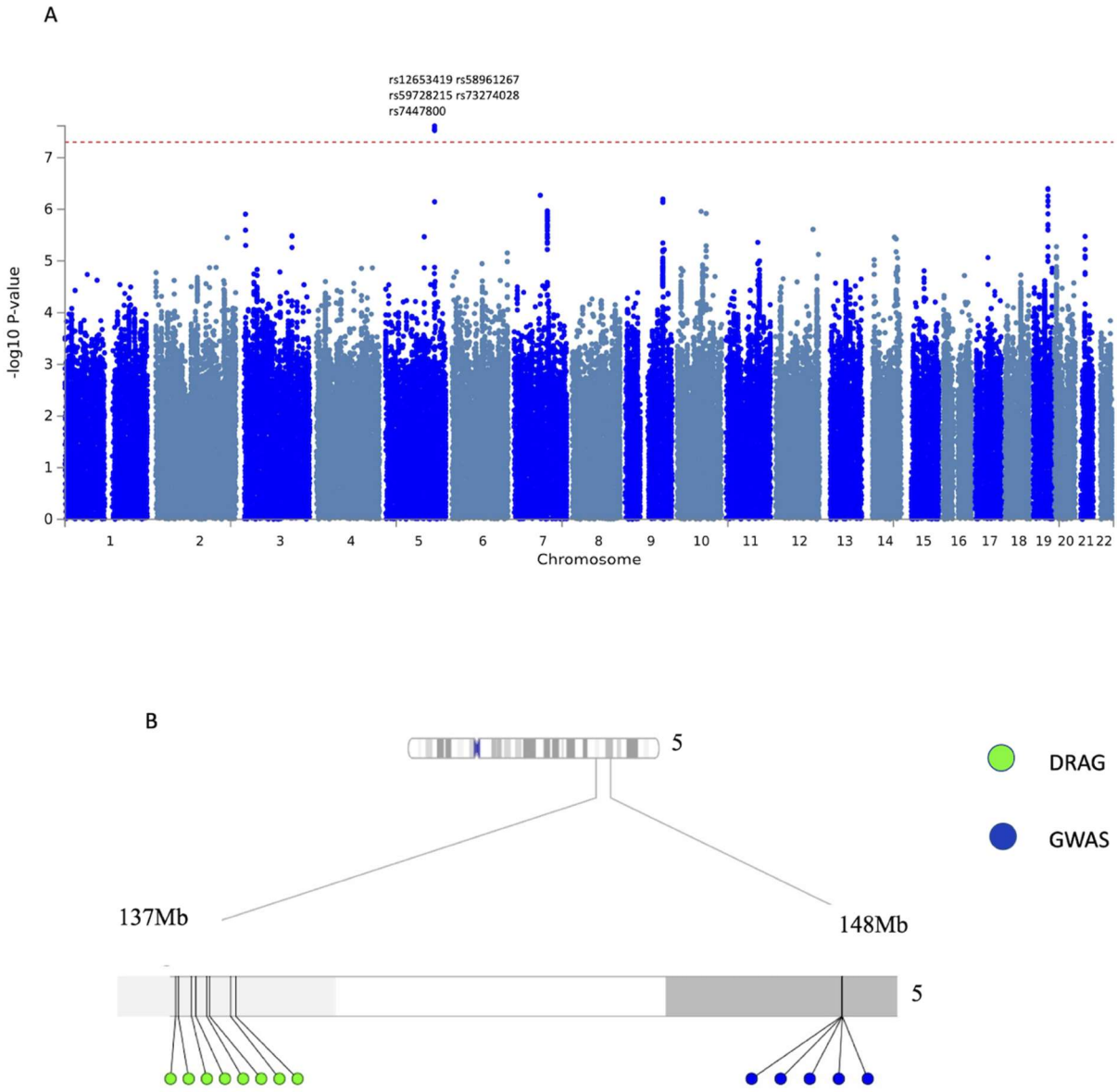


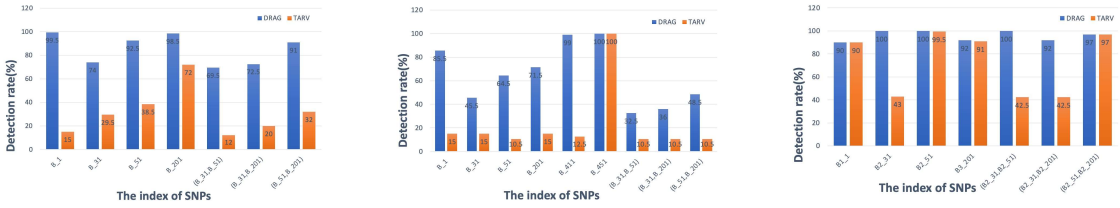
Figure 2: (**A**) Manhattan plot of traditional single SNP association analysis based on samples with white British ancestry only and controlled for gender, age and 10 PCs. The red dashed horizontal line corresponds to the commonly adopted genome-wide significant level at 5×10^{-8}. (**B**) The location of SNPs on chromosome 5 identified by DRAG (green) and traditional GWAS (blue).

DRAG increases detection power compared to tree-based model on simulated data

Finally, we perform simulation studies to demonstrate the efficacy of our approach, by comparing it to an established method, tree-based analysis of rare variants (TARV), in various settings. We refer the readers to [20] for more details about TARV implementation. The same procedure for both methods is repeated 200 times in each setting, and the pool of 200 results is summarized in Fig. 3. We evaluate the performance by detection rate, which is defined as the number of correct detections divided by the total number of repeated experiments.

Fig 3.A and Fig 3.B exhibit the comparisons between DRAG and TARV in single SNP detection and interacting SNPs detection, respectively. DRAG achieves higher detection rates than TARV in both single SNP detection and interacting SNPs detection. For example, the DRAG on SNP B_411 achieves detection rate of 99% for Case 1, which is better than the TARV on the same SNP (i.e.,12.5%). In addition, DRAG for interacting terms (B_31, B_201) has a detection rate of 32.5% for Case 1 and 69.5% for Case 2, while the TARV only achieves detection rate of 10.5% and 12% respectively, which suggests that DRAG performs better on detecting interactions. Overall, the DRAG outperforms TARV in detecting interacting terms.

Fig 3.C summarizes the performance of DRAG and TARV in the third setting. Both DRAG and TARV work well on detecting individual SNP B1_1 and interactions (B2_51, B3_201), as evidenced by respective similar detection rate for DRAG and TARV of 97% on B1_1 and 97% on (B2_51, B3_201). It is noteworthy that DRAG generates similar results when detecting single SNP B1_1, B2_31 and B2_51 as TARV. However, DRAG on single SNP B2_31 achieves detection rate of 100%, which is significantly superior to the TARV (i.e., 43%). Similarly, as with the DRAG, the observed detection rate for interaction term of (B2_31, B2_51) and (B2_31, B3_201) (100% and 92%, respectively), are superior to TARV from (B2_31, B2_51) with 42.5% and (B2_31, B3_201) with 42.5%. The results suggest DRAG outperforms the TARV by a large margin, which implies that deep learning-based model is superior to tree-based model at detecting complex interacting SNPs when large amounts of data are presented.



(A) Case 1 **(B) Case 2** **(C) Case 3**

Figure 3. Detection rate of true signal and interaction in different cases. The x-axis of panels (A) through (C) represents the index of SNP. The y-axis of panels (A) through (C) represents the detection rate (%). Blue and red bars represent DRAG and TARV, respectively. The B_k represents the *k*th SNP for first two cases and the Bj_K represents the *k*th SNP in the SNP set *j* for the third case, $1 \leq i \leq 1{,}0000$, $1 \leq j \leq 30$, $1 \leq k \leq 500$.

Discussion

In summary, we train deep learning to learn the association between genetics variants and COVID-19 related mortality using genotyping data from the UK-biobank white British cohort. We identify 15 super variants and explored their relationships to COVID-19 related mortality. We also report 54 genes related to detected super variants, many of these genetic variants overlap with previously reported associations with lung-related phenotypes or COVID-19 outcomes or inflammatory diseases.

In addition to novel genetic discoveries in COVID-19, our approach is also novel in several statistical aspects and possesses unique strengths: the ranking and aggregation strategy by an iterative super variant search enables us to capture complex non-linear interacting SNPs and consequently and obtain group of interacting SNPs with high COVID-19 related mortality risk potential. In contrast to the classic CAE approach, the DRAG incorporates two innovative processes: iteratively running CAE procedure and model selection through BIC. These steps enable the user to explore various combinations of SNPs and select the optimal one. Finally, we demonstrate in simulation study that DRAG is superior to tree-based model by a significant margin when we detect complex interacting SNPs that are entangled in high-dimensional nonlinear representations. We believe this is a significant advantage of our strategy since host genetic factors that account for disease risk frequently include multiple complex interactions. Together, our findings provide timely

clues and prospective directions for better understanding the molecular pathogenesis of COVID-19, which may have implications for clinical treatment of this disease.

Limitations

Our study is subject to several limitations. The study population largely consisted of white British ancestry UK Biobank participants, limiting generalizability to other populations. Because only about 5% of the UK Biobank population had COVID-19 data available, it will be interesting to explore our findings when additional data and populations of diverse ancestry become available. Although one can vary the degrees of non-linearity and flexibility in the model by tunning hyper-parameters like K_0, learning rate and the number of layers, etc., the computation cost of DRAG would increase dramatically as the value of K_0 increases. We expect to implement parallel computation in future work to make our procedure more user-friendly. Finally, the roles of the identified super variants and associated genes are not substantiated by functional validation. Still, our findings warrant future investigation to learn the associations between genetic variants and the COVID-19 outcomes.

Method

Data and preprocessing

Cohort dataset. The data used in the preparation of this study were obtained from the UK Biobank (project ID: 42009) [54, 55]. The UKB was launched in 2006, led by Heart Foundation Professor Sir R. Collins. The primary goal of UKB has been to help researchers to better understand a range of common and life-threatening diseases. The UKB cohort recruited 500,000 individuals aged 40–69 in the UK via mailer from 2006 to 2010. In total, we analyze 148,654 participants with COVID-19 data (Field ID: 40100) as of late December 2022. Out of these participants, 28,652 (19.27%) had the positive polymerase chain reaction (PCR) test result.

Data preprocessing. We first discard the UKB samples for which age or sex was missing. Second, we remove samples identified as outliers in heterozygosity and missing rates. We then delete samples which were identified as putatively carrying sex chromosome configurations that are not either XX or XY. These were identified by looking at average log2Ratios for Y and X chromosomes. Next, we remove the participant who are excluded from kinship inference process. Finally, we only consider white British participants to limit the potential effect of population structure. After standard sample quality controls, there remain 28,097 of COVID-19 infected participants in our final analyses, of which 1,656 are deaths (5.89%) and 26441 are survivors (Fig 4.A).

Definition of Phenotype. To detect the genetic variants associated with the risk of COVID-19 related mortality, we define a death as a person who died due to COVID-19 infection and a survivor as a person who had the positive polymerase chain reaction (PCR) test result but survived.

Demographic variables. Age and sex (Field ID: 31, 34) are used as confounding variables in regression analyses to eliminate potential bias when testing the significance of association between super variants and the phenotype (Fig 4 B&C). For sex, we used the genetic sex when available, and the self-reported sex when genetic sex was not available.

Genotype data and quality control. We use imputed genotyping data (Field ID: 22801-22822) from UK Biobank as genetic predictors in the present study. SNPs with duplicated names and positions are deleted. In addition, we remove variants with minor allele frequency (MAF) ≤ 0.05 and disrupted Hardy-Weinberg equilibrium (p value $< 10^{-6}$). We retain in total 8,238,098 SNPs. We next divide the whole SNP dataset into 2734 non-overlapping local SNP-sets according to the physical position so that each SNP-set consists of SNPs within a segment of physical length 1Mbp (Fig 4.D).

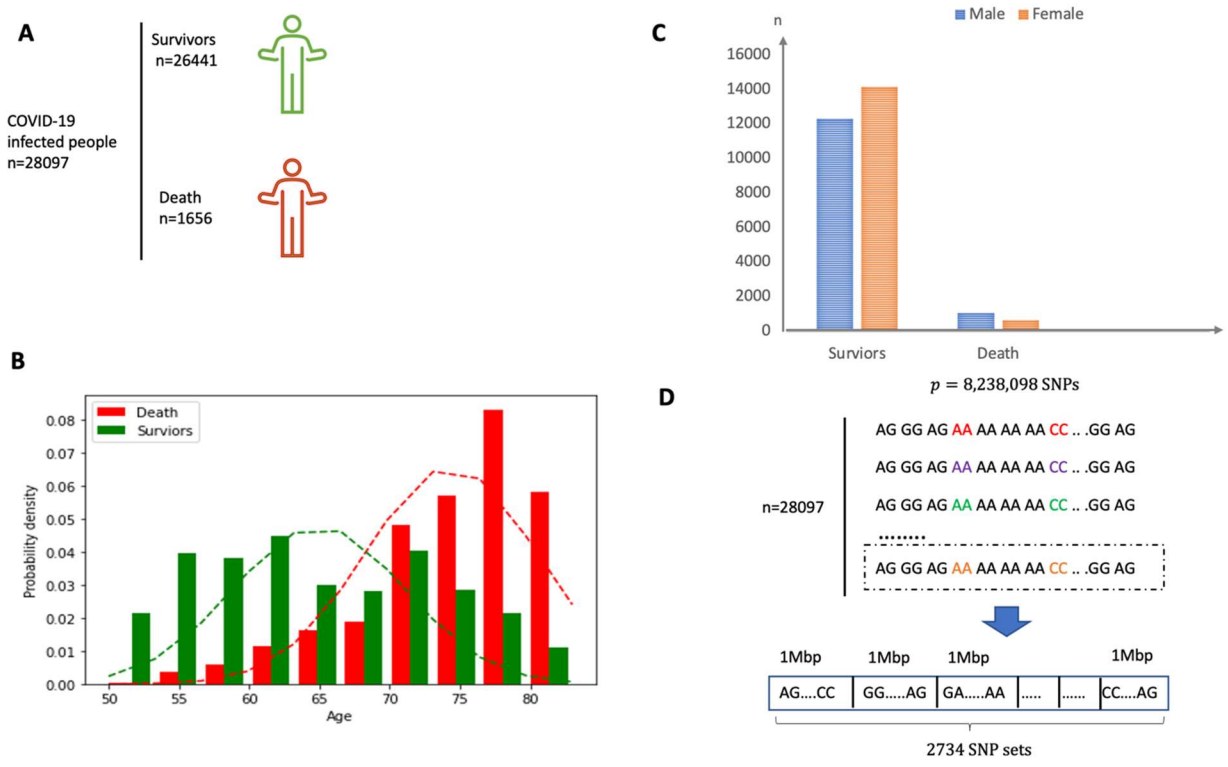


Figure 4: **(A)** Overview of the participants included and the samples and data collected. **(B)** Sex distribution in both survivor and death group. **(C)** Age distribution in both survivor and death group. The mean of age for death group is around 75 years old. **(D)** The SNP dataset are divided into 2734 non-overlapping local sets according to the physical position and each set consists of SNPs within a segment of physical length 1 Mbp.

Study design

All analyses are conducted in the UK Biobank unless otherwise stated. We first adopt the concept of super variants. In contrast to a gene, which refers to a physically connected region of a chromosome, the loci contributing to a super variant can locate in any part of the genome. Since the super variant aggregates the strength of distinct signals, it has been shown that it is both powerful and stable in association studies [17, 18]. Furthermore, super variants take into consideration the potential complex interactions among loci.

Our design considers the following discovery-validation procedure (Fig. 5). The whole dataset is partitioned into two sets, one for discovery and the other for verification with ratio of 2:1. A total of 1,104 fatalities and 17,627 survivors are included in the discovery set, whereas 8,814 deaths and 552 survivors are included in the verification set. We train the

DRGA to find the initial candidate super variants in the first half of discovery set, the logistic regression is then applied to the second half of the data from discovery set to find initial optimal super variants, permitting us control for Type I error. The SNPs contributing to the initial optimal super variants are extracted and aggregated into super variants on the verification datasets. We validate their associations with the COVID-19 related mortality through logistic regression on the verification set. We use 1.83×10^{-5} (i.e., 0.05/2734, adjusted for the number of super variants) as the Bonferroni-corrected cut-off p-value for one super variant on the discovery dataset. A super variant is verified if its logistic regression coefficient achieves the level of 0.05 significance on the verification dataset. Statistical analyses were conducted with Python 3.9.0.

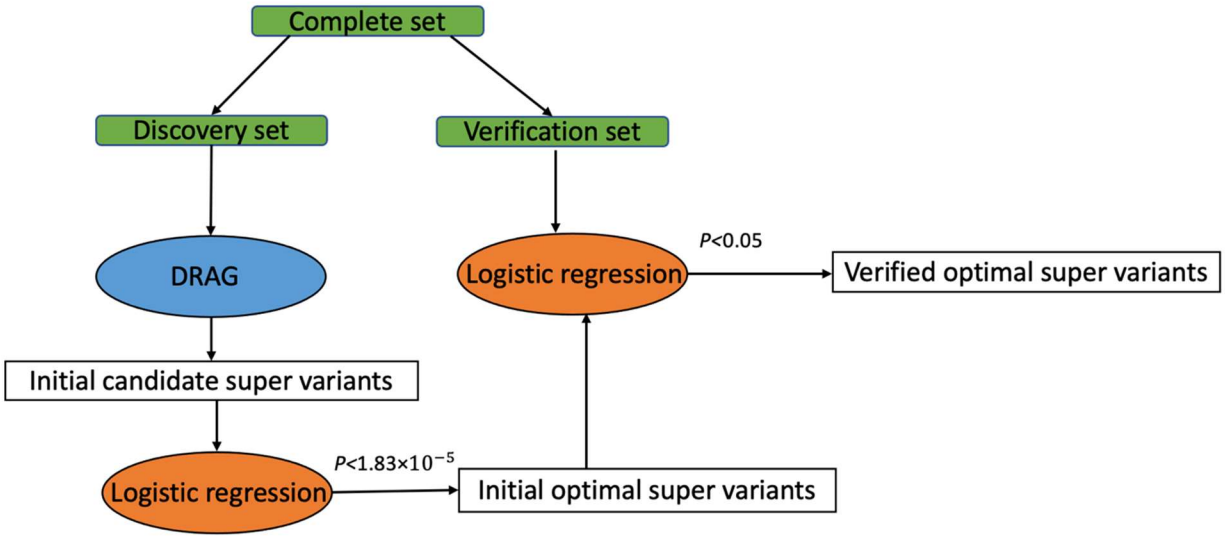


Figure 5. Overview of discovery-validation procedure. Complete dataset is partitioned into two sets, one for discovery and the other for verification. Discovery set is used to find optimal super variants and verification set is used for evaluating the selected super variants.

Machine learning algorithms

Overview of DRAG algorithm. DRAG algorithm consists of three main steps as summarized in Fig 6. In the first step, chromosomes are divided into different non-overlapping local genomic sets according to their physical positions. In the second step, the CAE is utilized within each set to select a subset of k SNPs, where k is user-specified and spans a certain

range, e.g., (1, K_0), K_0 is a hyper-parameter. For each chosen k, we train a CAE model. Later, these CAE models will be evaluated based on some criteria, and a best k will be picked. In another words, K_0 different subsets of SNPs are selected for each genomic set by iteratively training CAE. In the third step, a super variant is formed by a subset of SNPs that indicates whether any minor allele is present [16]. A total of K_0 super variants are formed. We use logistic regression combined with BIC to identify the optimal super variant among the K_0 super variants with each genomic set. The optimal super variant is the set of SNPs that minimizes the BIC value.

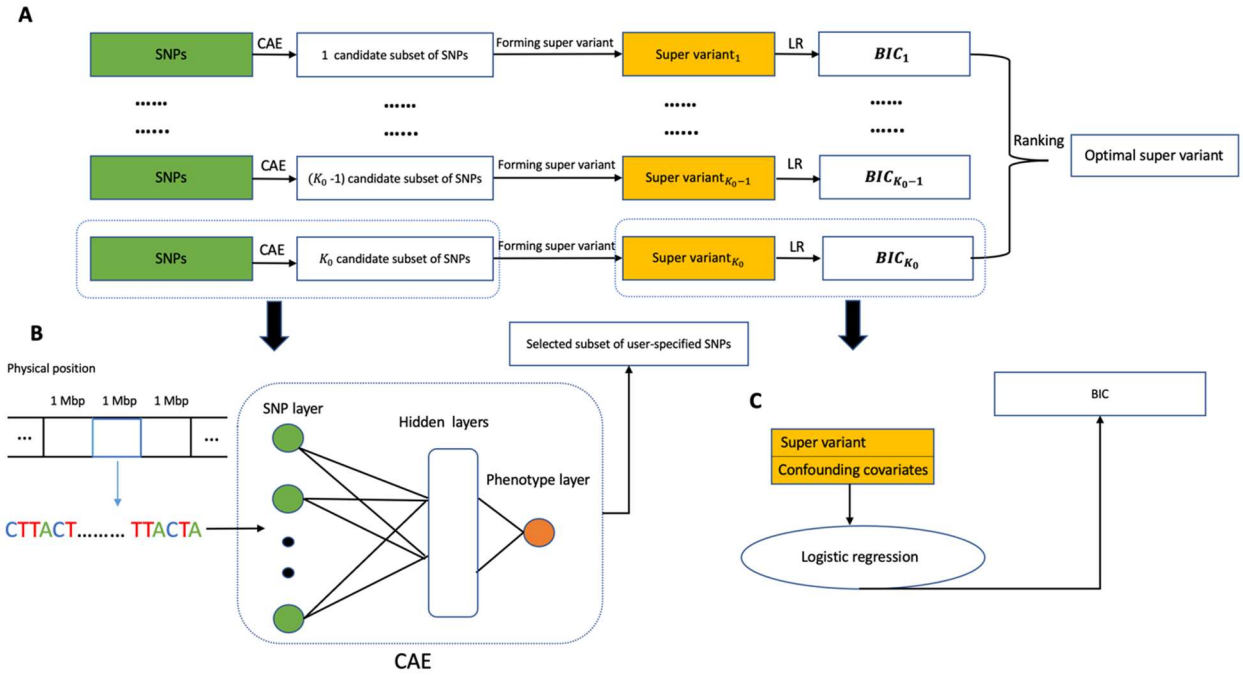


Figure 6: Visual representation of the proposed SNP selection approach using DRAG. **(A)** An overall flowchart of the approach for selecting of super variant. **(B)** CAE model training and finding the selected subset of user-specified SNPs. **(C)** Logistic regression training and finding BIC value.

Concrete Auto-encoder architectures. We design a deep neural network architecture to learn complex relations between genetic variants and phenotypes (Fig 7.A). Let vector $\boldsymbol{x} \in R^p$ be a collection of SNPs and y be a phenotype which is a binary random variable taking value 0 or 1. We first use a user-specified set of k SNPs denoted by $\boldsymbol{z}$. We have the $\boldsymbol{z} = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{x}$, where $\boldsymbol{z} \in R^k$ and $\boldsymbol{B} \equiv [\boldsymbol{\beta}_1, .., \boldsymbol{\beta}_k] \in R^{p\times k}$ denotes the parameters between SNPs layer and user-specified subset layer. Thus, $z_i = \boldsymbol{\beta}_i{}^T x = x_1\,\beta_{i1} + ... + x_p\beta_{ip}$, $i = 1, \ldots, k$, is referred to as to the

ith nodes in user-specified subset layer. Decoder layers consist of hidden layers and phenotype layers and serve as the reconstruction function. Decoder layers take the outputs z from user-specified subset as its inputs. A decoder with h hidden layers is defined as $f(z) = f^{(h)}(...f^2(f^1(z)))$, where $f^{(i)}(\cdot)$ represents the ith hidden layers and the output $f(z)$ consists of the predicted phenotype for an individual. The loss function for binary classification is given by:

$$l = -y \log f(z) - (1 - y)(1 - \log f(z)) \quad (1)$$

In our specific application, we let $h = 2$, n_1 and n_2 denote the number of nodes in each layer, and $\Gamma_1 \in R^{n_1 \times k}$ $\Gamma_2 \in R^{n_1 \times n_2}$, $\boldsymbol{b}_1 \in R^{n_1}$ and $\boldsymbol{b}_2 \in R^{n_2}$ denotes the weights and bias term attached to each hidden layer j (j = 1,2). Then the outputs in each hidden layer are $f^1(z) = \phi_1(\Gamma_1 z + \boldsymbol{b}_1)$ and $f^2(z) = \phi_2(\Gamma_2 f^1(z) + \boldsymbol{b}_1)$, where ϕ_j are activation functions of each layer. In particular, the activation functions for the hidden layers are ReLU: $\phi(x) = \max(0, x)$. The predicted $\tilde{y}$ can be written as $f(z) \equiv \tilde{y} = \phi_3(\Gamma_3 f^2(z) + b_3)$, where ϕ_3 is the activation function, which is typically a logistic function for dichotomous traits, $\Gamma_3 \in R^{n2}$ and b_3 are the weights and bias in phenotype layer. Therefore, the optimization function for the data under the truth can be written as:

$$R(f) = -\frac{1}{n}\sum_{j=1}^{n} y_i \log f(\boldsymbol{x_j}) - (1 - y_j)(1 - \log f(\boldsymbol{x_j})) \quad (2)$$

where n is the total number of subjects and x_i and y_i are set of SNPs and phenotype of ith subject, respectively.

In practice, the user-specified subset layer generally outputs weighted linear combination of input variants. We next describe how the user-specified subset layers works, that is, how each variant is selected (Fig 7.B). To preform the variant selection, the element of $\boldsymbol{\beta}$ has be to either 0 or 1. We now only consider ith node in user-specified subset layer and sample p-dimensions vector $\boldsymbol{\beta}_i$ using the proposed technique in [23, 56, 57]:

$$\beta_{ij} = \frac{\exp((\log\alpha_j + g_j)/T)}{\sum_{t=1}^{p} \exp((\log\alpha_t + g_t)/T)} \quad (3)$$

where β_{ij} refers to the jth element in $\boldsymbol{\beta}_i$, $T \in (0, \infty)$ is the so-called temperature parameter, $\alpha \in R^p$ is a p-dimensional vector which all the elements are strictly greater than zero and $g \in R^p$ is a p-dimensional vector which all elements are generated from

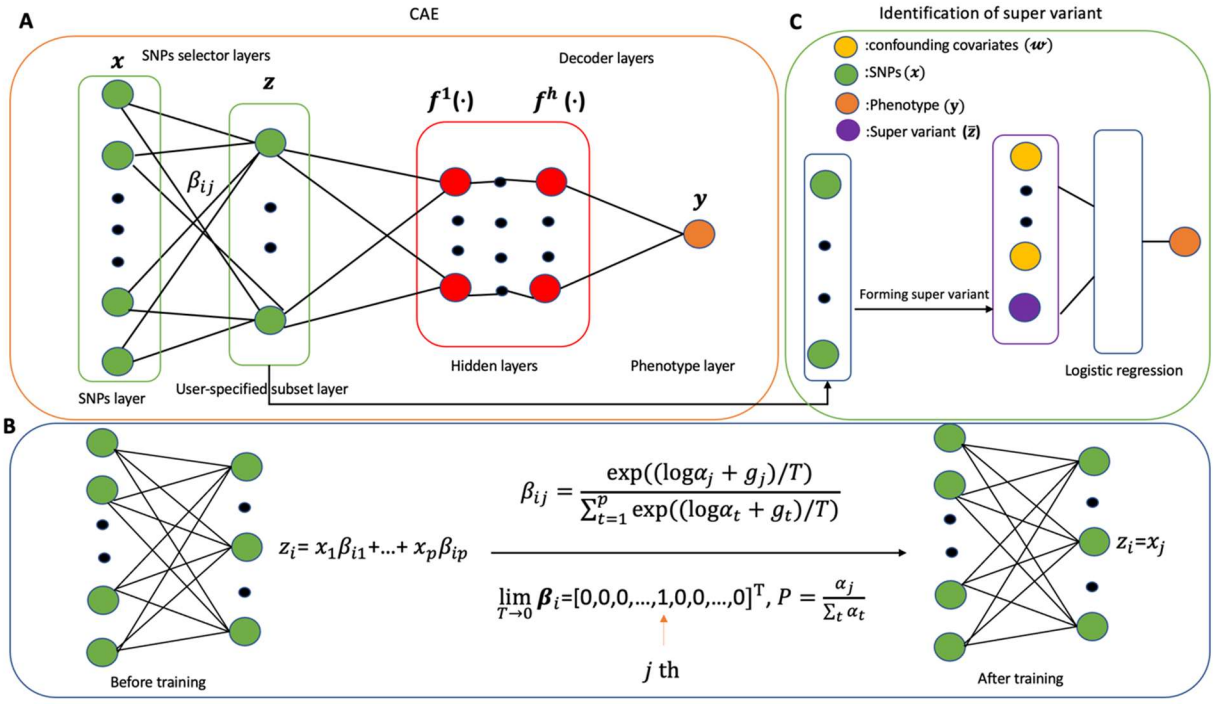


Figure 7: (A) An overall of representation of the Concrete auto-encoder approach. x and y are the set of SNPs and the phenotype, respectively. z denotes the set of top identified SNPs and $f^{(i)}(\cdot)$ represents the outputs in each hidden layers. **(B)** A mathematical overview of SNP selection procedure using CAE on SNPs selector layers. **(C)** The identified SNPs from (A) and confounding covariates are then used for logistic regression to facilitate the identification of super variants.

a Gumbel distribution [58]. In this way, we obtain the $\boldsymbol{\beta}_i$ smoothly approaches the discrete distribution as $T \to 0$, outputting one hot vectors with $\beta_{ij} = 1$ with probability $\frac{\alpha_j}{\sum_t \alpha_t}$. That is $\lim_{T\to 0} \beta_i = [0,0, \ldots 1,0,0, \ldots,0]^{\mathrm{T}}$. Therefore, the ith node in user-specified subset layer outputs exactly one of the variants when $T \to 0$. Using the same logic we sample a p-dimensional random vector $\boldsymbol{\beta}$ for each of the k nodes to perform variant selection. Note, for the initialization of T, Abid [23] claimed that the model is unable to explore different feature combinations and converges to a poor local minimum if the initial value of T is set to be extremely low. We apply their work and use the varying temperature $T(e) = T_1(T_2/T_1)^{e/E}$, where $T(e)$ is the temperature at epoch number e, and E is the total number of epochs to train the model, T_1 and T_2 are tuning parameters which are usually set to be a high value for T_1 and

a low value for T_2. With such a choice of $\boldsymbol{\beta}$, the optimization function (2) can be minimized with respected to $\boldsymbol{\alpha}$,$\boldsymbol{\Gamma}_j$ and $\boldsymbol{b}_j$ (j = 1,2,3) using the backpropagation algorithm and the user-specified set of k SNPs are selected afterwards.

Identification of super variants through logistic regression. Since the optimal number of SNPs remains unknow, the super variants learned by CAE may be sub-optimal. To resolve this issue, we use a model selection technique with BIC in logistic regression, in which the target is the phenotype and the main effect estimated by super variant that consists of selected SNPs (Fig 7.C). To facilitate the identification of super variants, all possible values for the number of nodes in user-specified set layer with observations z are inspected using logistic regression. The range of values to consider for the k is greater than 1 and less than K_0, where $K_0 \leq p$ and p is the total number of variants. For each candidate value $k \in [1, K_0]$, the model with smallest BIC value is selected and the corresponding variants aggregated to form the optimal super variants. As an illustration, let $z_1,...z_k$ be the variants selected by CAE and $\overline{z_k}$ denote super variant and w_{iq} denote the qth covariate of subject i. A general model is given by

$$g(\mu_i) = \sum_q \gamma_q w_{iq} + \omega \overline{z_{ik}} \quad (4)$$

where $\mu_i = E(y_i)$, y_i is the phenotype for subject i, and g is a link function which is logit function for binary classification, $\overline{z_k}$ =I $(\sum_{r=1}^{k} z_r > 0)$, and γ_q and ω are coefficients for covariates and genetic variants, respectively. We next apply model (4) on formed super variant and confounding variables to find the BIC value denoted by BIC . The importance of each model can be evaluated based on the magnitude of BIC. Finally, we choose the super variant which consists of k_b variants $(z_1,...,z_{k_b})$ with the smallest BIC as the optimal super variant, where $\{k_b : BIC_{k_b} \leq BIC_i, i = 1,.., k_{b-1},k_{b+1},...,K_0\}$

Tuning. There is a negligible impact of tunning parameters, K_0 , learning rate, the number of layers and the number of nodes in each layer on model performance. Especially, the value of K_0 is critical for the success of our algorithm. The range of values to consider for the K_0

for each set is no more than the total number of SNPs in this set. Too large a K_0 would result in unnecessary computation burden, while too small a K_0 might produce a sub-optimal super variant. In this study, we choose K_0 to be 25 and use 128 and 64 nodes to each hidden layer of 2-hidden layers CAE with the Adam optimizer of learning rate of 0.001, which is tuned on an independent identically distributed dataset.

Simulation framework

For the simulated data set up, we use the genotype data on Chromosome 22 from the UK Biobank COVID-19 dataset because there is no significant signal identified on this chromosome in previous studies. We choose 10,000 patients at random. For the first two cases, we randomly sample 500 SNPs from a single SNP set to form the synthetic genetic dataset. In the third case, we randomly select 30 SNP sets and 500 random SNPs from each set to create the whole synthetic genetic dataset. As a consequence, a super variant discovered by both DRAG and TARV is considered to be significant if the p-value is less than 0.05 (Case 1 and 2) and 0.05/30 (Case 3), respectively. The disease y is generated from $y_i \sim$ Bernoulli (p_i), and

Case 1:$\log\frac{p_i}{1-p_i} = 0.99[I((x_{i\_B1} + x_{i\_B31}x_{i\_B5} + x_{i\_B3}\ x_{i\_B201} + x_{i\_B5}\ x_{i\_B201}) > 0) + I(x_{i\_B411} > 0) + I(x_{i\_B45} > 0)]$

Case 2:$\log\frac{p_i}{1-p_i} = 0.99[I((x_{i\_B1} + x_{i\_B3}\ x_{i\_B51}) > 0) + I((x_{i\_B1} + x_{i\_B31}x_{i\_B201}) > 0) + \mathrm{I}(x_{i\_B51} > 0) + (x_{i\_B201} > 0)]$

Case 3:$\log\frac{p_i}{1-p_i} = 0.99[I((x_{i\_B1\_1} + x_{i\_B2\_31}x_{i\_B2\_51}) > 0) + I((x_{i\_B1\_1} + x_{i\_B2\_31}x_{i\_B3\_201}) > 0) + \mathrm{I}(x_{i\_B2\_51} > 0) + (x_{i\_B3\_201} > 0)]$

where x_{i_Bk} is the kth SNP for subject i for first two cases and x_{i_Bjk} represents the kth SNP in SNP set j for subject i for the last case, $1 \le i \le 1{,}0000$, $1 \le j \le 30$, $1 \le k \le 500$. Case 1 includes individual effect expressed as B_411 and B_451. In addition, both cases consist of same interaction term of (B_31, B_201) a (B_31, B_51) and (B_51, B_201). There are a total of 6 true signals in Case 1 and 4 true signals in Case 2. We consider an interaction term being identified if all true signal SNPs within the same set, such as the two SNPs B_31 and B_51 are identified for case 1 at the same time. For brevity, Case 3 consists of 4 true signals

from three different SNP sets with two different structures, individual signal, interacting signal with group sizes 2.

Data availability

UK Biobank data is made available to researchers from universities and other research institutions with genuine research inquiries, following the UK Biobank approval. (https://www.ukbiobank.ac.uk).

Author contributions

HZ conceived and oversaw this project, ZL and WD took the main responsibility in execution of this project, SW and YY contributed to the data analysis. All authors made critical input to the manuscript.

Acknowledgments

Zhang's research is supported in part by U.S. National Institutes of Health (R01HG010171 and R01MH116527) and National Science Foundation (DMS-2112711). This research has been conducted using the UK Biobank Resource under Application Number 42009. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure.

References

1. Khalili, M., et al., *Epidemiological characteristics of COVID-19: a systematic review and meta-analysis.* Epidemiology & Infection, 2020. **148**.

2. Lao, X., et al., *The epidemiological characteristics and effectiveness of countermeasures to contain coronavirus disease 2019 in Ningbo City, Zhejiang Province, China.* Scientific reports, 2021. **11**(1): p. 1-12.
3. Khailany, R.A., M. Safdar, and M. Ozaslan, *Genomic characterization of a novel SARS-CoV-2.* Gene reports, 2020. **19**: p. 100682.
4. D'Antonio, M., et al., *SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affecting gene expression in a variety of tissues.* Cell reports, 2021. **37**(7): p. 110020.
5. Docherty, A.B., et al., *Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study.* bmj, 2020. **369**.
6. Stoian, A.P., et al., *Gender differences in the battle against COVID-19: impact of genetics, comorbidities, inflammation and lifestyle on differences in outcomes.* International journal of clinical practice, 2020.
7. Sharma, G., A.S. Volgman, and E.D. Michos, *Sex differences in mortality from COVID-19 pandemic: are men vulnerable and women protected?* Case Reports, 2020. **2**(9): p. 1407-1410.
8. Jin, J.-M., et al., *Gender differences in patients with COVID-19: focus on severity and mortality.* Frontiers in public health, 2020: p. 152.
9. Pareek, M., et al., *Ethnicity and COVID-19: an urgent public health research priority.* The Lancet, 2020. **395**(10234): p. 1421-1422.
10. Group, S.C.-G., *Genomewide association study of severe Covid-19 with respiratory failure.* New England Journal of Medicine, 2020. **383**(16): p. 1522-1534.
11. Shelton, J.F., et al., *Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity.* Nature Genetics, 2021. **53**(6): p. 801-808.
12. Pairo-Castineira, E., et al., *Genetic mechanisms of critical illness in Covid-19.* Nature, 2021. **591**(7848): p. 92-98.
13. Initiative, C.-H.G., *The COVID-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic.* European Journal of Human Genetics, 2020. **28**(6): p. 715.
14. Schmiedel, B.J., et al., *COVID-19 genetic risk variants are associated with expression of multiple genes in diverse immune cell types.* Nature communications, 2021. **12**(1): p. 1-12.
15. Hu, J., et al., *Supervariants identification for breast cancer.* Genetic epidemiology, 2020. **44**(8): p. 934-947.
16. Song, C. and H. Zhang, *TARV: Tree-based analysis of rare variants identifying risk modifying variants in CTNNA2 and CNTNAP2 for alcohol addiction.* Genetic epidemiology, 2014. **38**(6): p. 552-559.
17. Hu, J., et al., *Genetic variants are identified to increase risk of COVID-19 related mortality from UK Biobank data.* Human genomics, 2021. **15**(1): p. 1-10.
18. Li, T., et al., *Super-variants identification for brain connectivity.* Human brain mapping, 2021. **42**(5): p. 1304-1312.
19. Eraslan, G., et al., *Deep learning: new computational modelling techniques for genomics.* Nature Reviews Genetics, 2019. **20**(7): p. 389-403.
20. Xiong, H.Y., et al., *The human splicing code reveals new insights into the genetic determinants of disease.* Science, 2015. **347**(6218): p. 1254806.

21. Arloth, J., et al., *DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning.* PLoS computational biology, 2020. **16**(2): p. e1007616.
22. Zou, Q., et al., *Deep learning based feature selection for remote sensing scene classification.* IEEE Geoscience and Remote Sensing Letters, 2015. **12**(11): p. 2321-2325.
23. Balın, M.F., A. Abid, and J. Zou. *Concrete autoencoders: Differentiable feature selection and reconstruction*. in *International conference on machine learning*. 2019. PMLR.
24. Niemi, M.E., et al., *Mapping the human genetic architecture of COVID-19.* Nature, 2021. **600**: p. 472-477.
25. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nature genetics, 2006. **38**(8): p. 904-909.
26. Marchini, J., et al., *The effects of human population structure on large genetic association studies.* Nature genetics, 2004. **36**(5): p. 512-517.
27. Watanabe, L.M., et al., *The influence of bitter-taste receptor (TAS2R) expression in pharmacological response to Chloroquine in obese patients with COVID-19*. 2020, SciELO Brasil.
28. Jansen, J., et al., *SARS-CoV-2 infects the human kidney and drives fibrosis in kidney organoids.* Cell stem cell, 2022. **29**(2): p. 217-231. e8.
29. Morsy, S., *NCAM protein and SARS-COV-2 surface proteins: In-silico hypothetical evidence for the immunopathogenesis of Guillain-Barré syndrome.* Medical hypotheses, 2020. **145**: p. 110342.
30. Laudanski, K., et al., *Guillain–Barré Syndrome in COVID-19—The Potential Role of NCAM-1 and Immunotherapy.* BioMed, 2021. **1**(1): p. 80-92.
31. Mastropasqua, L., et al., *Transcriptomic analysis revealed increased expression of genes involved in keratinization in the tears of COVID-19 patients.* Scientific Reports, 2021. **11**(1): p. 1-12.
32. Gao, W., et al., *Identification of NCAM that interacts with the PHE-CoV spike protein.* Virology Journal, 2010. **7**(1): p. 1-11.
33. Li, Z., et al., *Gene expression profiling in autoimmune noninfectious uveitis disease.* The Journal of Immunology, 2008. **181**(7): p. 5147-5157.
34. Upadhyaya, J., et al., *Bitter taste receptor T2R1 is activated by dipeptides and tripeptides.* Biochemical and biophysical research communications, 2010. **398**(2): p. 331-335.
35. Barham, H.P., et al., *Association Between Bitter Taste Receptor Phenotype and Clinical Outcomes Among Patients With COVID-19.* JAMA network open, 2021. **4**(5): p. e2111410-e2111410.
36. Gustafson, E.A. and G.M. Wessel, *DEAD-box helicases: posttranslational regulation and function.* Biochemical and biophysical research communications, 2010. **395**(1): p. 1.
37. Oshiumi, H., et al., *DDX60 is involved in RIG-I-dependent and independent antiviral responses, and its function is attenuated by virus-induced EGFR activation.* Cell reports, 2015. **11**(8): p. 1193-1207.
38. Byun, J., et al., *Genome-wide association study of familial lung cancer.* Carcinogenesis, 2018. **39**(9): p. 1135-1140.

39. Squeglia, F., et al., *Host DDX helicases as possible SARS-CoV-2 proviral factors: a structural overview of their hijacking through multiple viral proteins.* Frontiers in chemistry, 2020. **8**: p. 1150.
40. Chen, T.H.-P., et al., *Knockdown of Hspa9, a del (5q31. 2) gene, results in a decrease in hematopoietic progenitors in mice.* Blood, The Journal of the American Society of Hematology, 2011. **117**(5): p. 1530-1539.
41. Krysiak, K., et al., *Reduced levels of Hspa9 attenuate Stat5 activation in mouse B cells.* Experimental hematology, 2015. **43**(4): p. 319-330. e10.
42. Tay, M.Z., et al., *Decreased memory B cells frequencies in COVID-19 Delta variant vaccine breakthrough infection.* EMBO molecular medicine, 2022: p. e15227.
43. Aveyard, P., et al., *Association between pre-existing respiratory disease and its treatment, and severe COVID-19: a population cohort study.* The lancet Respiratory medicine, 2021. **9**(8): p. 909-923.
44. Melms, J.C., et al., *A molecular single-cell lung atlas of lethal COVID-19.* Nature, 2021. **595**(7865): p. 114-119.
45. Wang, S., et al., *A single-cell transcriptomic landscape of the lungs of patients with COVID-19.* Nature cell biology, 2021. **23**(12): p. 1314-1328.
46. Kichaev, G., et al., *Leveraging polygenic functional enrichment to improve GWAS power.* The American Journal of Human Genetics, 2019. **104**(1): p. 65-75.
47. Fumagalli, A., et al., *Long-term changes in pulmonary function among patients surviving to COVID-19 pneumonia.* Infection, 2021: p. 1-4.
48. Lewis, K.L., et al., *COVID-19 and the effects on pulmonary function following infection: A retrospective analysis.* EClinicalMedicine, 2021. **39**: p. 101079.
49. Zhang, S., et al., *Eight months follow-up study on pulmonary function, lung radiographic, and related physiological characteristics in COVID-19 survivors.* Scientific reports, 2021. **11**(1): p. 1-13.
50. Gomez, G.N., et al., *SARS coronavirus protein nsp1 disrupts localization of Nup93 from the nuclear pore complex.* Biochemistry and Cell Biology, 2019. **97**(6): p. 758-766.
51. Slavin, T.P., et al., *Two-marker association tests yield new disease associations for coronary artery disease and hypertension.* Human genetics, 2011. **130**(6): p. 725-733.
52. Leung, A.K., et al., *The Conserved Macrodomain Is a Potential Therapeutic Target for Coronaviruses and Alphaviruses.* Pathogens, 2022. **11**(1): p. 94.
53. Michalska, K., et al., *Crystal structures of SARS-CoV-2 ADP-ribose phosphatase: from the apo form to ligand complexes.* IUCrJ, 2020. **7**(5): p. 814-824.
54. Armstrong, J., et al., *Dynamic linkage of covid-19 test results between public health england's second generation surveillance system and uk biobank.* Microbial genomics, 2020. **6**(7).
55. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS medicine, 2015. **12**(3): p. e1001779.
56. Maddison, C.J., A. Mnih, and Y.W. Teh, *The concrete distribution: A continuous relaxation of discrete random variables.* arXiv preprint arXiv:1611.00712, 2016.
57. Jang, E., S. Gu, and B. Poole, *Categorical reparameterization with gumbel-softmax.* arXiv preprint arXiv:1611.01144, 2016.
58. Gumbel, E.J., *Statistical theory of extreme values and some practical applications: a series of lectures*. Vol. 33. 1954: US Government Printing Office.

Supplementary information

Super Variants	SNP	Position	Minor allele	Major allele	OR	p-value	Nearest gene	Prior
chr2_35	rs62130610	34226429	C	A	1.184	4.41×10^{-2}	LINC01317	
	rs10495797	34117511	G	A	1.069	1.30×10^{-1}	LINC01317	
	rs570706090	34981873	T	TA	1.129	1.97×10^{-3}	LINC01320	[1]
chr4_24	rs4631028	23868735	T	A	1.327	5.35×10^{-3}	PPARGC1A	
	rs6839807	23781974	T	C	1.086	1.79×10^{-1}	GBA3	
	rs75544795	23461557	C	T	1.105	1.59×10^{-1}	GBA3	
	rs10517008	23272197	C	G	1.307	9.86×10^{-5}	GBA3	
chr4_170	4:169754442_CCCACCCTGCGGAG_C	169754442	C	CCCACCCTGCGGAG	1.037	7.24×10^{-1}	PALLD	
	4:169326163_AT_A	169326163	A	AT	1.123	1.01×10^{-2}	DDX60L	
	4:169513881_TA_T	169513881	T	TA	1.145	2.00×10^{-3}	PALLD	
	rs6553042	169636877	A	G	1.111	1.10×10^{-1}	PALLD	
chr5_10	rs12517344	9455377	T	A	1.143	2.52×10^{-1}	SEMA5A	
	rs72733036	9735819	G	A	1.169	7.99×10^{-4}	LINC02112	
	rs190052994	9001770	C	A	1.354	1.88×10^{-2}	LINC02199	
	rs34723029	9189317	T	C	1.071	3.51×10^{-1}	SEMA5A	
	rs72734818	9838544	T	C	1.131	6.95×10^{-2}	LINC02112	
	5:9305797_GTA_G	9305797	G	GTA	1.309	8.98×10^{-3}	SEMA5A	
	rs180899355	9584553	C	T	1.343	1.13×10^{-3}	TAS2R1	[1]
ch5_138	rs200095891	137588969	GCC	G	1.383	3.04×10^{-2}	GFRA3	
	rs17228325	137546642	A	G	1.086	3.16×10^{-1}	CDC23	
	rs62381756	137315567	A	T	1.129	1.58×10^{-1}	FAM13B	
	rs41294550	137911213	C	G	1.178	1.06×10^{-1}	HSPA9	
	rs78295109	137990043	T	C	0.959	6.41×10^{-1}	HSPA9	
	rs61537055	137076071	T	C	1.163	1.34×10^{-1}	KLHL3	
	rs146667750	137116264	CAA	C	1.047	5.31×10^{-1}	HNRNPA0	
	rs539251420	137388271	A	C	1.206	5.94×10^{-2}	LOC100130172	
ch6_54	6:53464104_GA_G	53464104	G	GA	1.136	7.69×10^{-3}	LOC101927136	
	rs143431923	53406014	CA	C	1.740	1.05×10^{-5}	GCLC	
	rs141615401	53247580	C	T	1.083	5.84×10^{-1}	ELOVL5	
	rs11967823	53016850	G	A	1.203	1.15×10^{-2}	GCM1	
	rs12197493	53236341	T	C	1.048	3.72×10^{-1}	ELOVL5	
ch6_163	rs77426763	162624682	C	T	1.083	1.44×10^{-1}	PRKN	
	rs112013442	162825837	A	G	1.133	9.61×10^{-2}	PRKN	
	rs13205685	162279763	A	G	1.054	1.99×10^{-1}	PRKN	
	rs9365309	162043005	A	G	1.121	2.60×10^{-2}	PRKN	
	rs73032254	162501137	C	A	1.085	1.78×10^{-1}	PRKN	

ch7_43	rs73096390	42282365	T	C	1.125	8.45×10^{-2}	GLI3	
	rs608045	42940405	G	A	1.061	1.39×10^{-1}	LINC01448	
	rs78659910	42182080	C	T	1.360	3.31×10^{-3}	GLI3	
	rs12539096	42492769	G	A	1.163	1.84×10^{-2}	GLI3	
	rs62457857	42773707	A	G	1.140	9.61×10^{-2}	LINC01448	
	rs112043733	42932691	A	G	1.144	5.14×10^{-2}	LINC01448	
ch7_151	rs76398985	150299940	G	C	1.060	2.38×10^{-1}	GIMAP4	
	rs6943608	150827207	A	G	1.219	4.80×10^{-6}	AGAP3	
	rs2052130	150548598	A	G	1.144	1.91×10^{-2}	TMEM176A	
	7:150989011_CT_C	150989011	CT	C	1.037	3.81×10^{-1}	SMARCD3	
	rs118033050	150586206	T	G	1.094	2.14×10^{-1}	AOC1	
	rs12540488	150895164	C	G	1.052	4.57×10^{-1}	IQCA1L	
ch10_6	10:5701622_AC_A	5701622	A	AC	1.065	1.74×10^{-1}	ASB13	
	rs3905460	5112190	G	A	1.347	1.55×10^{-2}	AKR1C3	
	10:5007398_GTA_G	5007398	G	GTA	1.231	6.18×10^{-2}	AKR1C1	
	rs61293789	5108313	T	G	1.213	4.72×10^{-4}	AKR1C3	
	rs7918112	5621174	G	A	1.172	1.42×10^{-2}	CALML3	
chr11_114	rs149645066	113585316	A	G	1.148	4.46×10^{-1}	TMPRSS5	
	rs117517757	113791826	A	C	1.122	9.82×10^{-2}	HTR3B	
	rs11214650	113408657	C	T	1.134	2.12×10^{-2}	DRD2	
	rs17115095	113060660	C	A	1.134	1.88×10^{-1}	NCAM1	[1]
	rs73000932	113954513	A	G	1.179	3.93×10^{-4}	ZBTB16	[2]
	rs4938008	113224127	T	C	1.049	7.21×10^{-1}	TTC12	
	rs56363150	113447787	A	G	0.715	5.48×10^{-3}	DRD2	
chr13_92	13:91800545_TA_T	91800545	T	TA	1.192	8.00×10^{-4}	LINC00379	
	rs72638966	91763846	C	G	1.258	1.44×10^{-1}	LINC00380	
	rs117152025	91991849	T	C	1.145	1.11×10^{-1}	LINC00379	
	rs9523036	91313809	T	C	1.034	5.86×10^{-1}	LINC01049	
	rs75068313	91477263	A	C	0.999	9.92×10^{-1}	LINC01049	
	rs77578051	91314988	T	C	1.183	1.31×10^{-2}	LINC01049	
	rs1325330	91048937	A	C	1.031	6.18×10^{-1}	MIR622	
	rs112350997	91839578	G	A	1.113	2.51×10^{-1}	LINC00379	
chr16_57	rs13306676	56921829	T	C	1.159	1.51×10^{-2}	SLC12A3	
	rs141590048	56648194	T	C	1.305	1.22×10^{-2}	MT2A	
	rs77356445	56600763	T	G	1.018	8.52×10^{-1}	MT4	
	rs146882751	56371699	T	C	0.881	4.52×10^{-1}	GNAO1	
	rs7206202	56084079	C	G	1.076	7.24×10^{-2}	CES5A	
	rs117501040	56238809	T	C	0.979	8.66×10^{-1}	GNAO1	
	rs75985606	56102938	T	C	1.075	3.62×10^{-1}	CES5A	

	rs12596894	56840027	C	T	1.110	5.70×10^{-2}	NUP93
	rs12444694	56920104	A	G	1.183	6.72×10^{-3}	SLC12A3
chr17_49	rs55649994	48355968	T	C	1.193	2.27×10^{-2}	TMEM92
	rs79806280	48005420	A	G	1.071	5.07×10^{-1}	TAC4
	rs35785126	48037287	G	C	1.218	2.54×10^{-3}	TAC4
	rs113706230	48281439	A	G	1.104	3.95×10^{-1}	COL1A1
	rs552888360	48685777	G	GT	1.321	1.32×10^{-2}	CACNA1G
chr20_15	rs8123267	14169136	C	T	1.126	1.59×10^{-1}	MACROD2
	rs73096611	14467744	T	C	1.177	2.13×10^{-1}	MACROD2
	20:14358660_AT_A	14358660	A	AT	1.042	6.79×10^{-1}	MACROD2
	rs73093934	14981772	A	T	1.141	3.74×10^{-1}	MACROD2
	rs79453392	14413334	A	G	1.194	1.21×10^{-1}	MACROD2
	rs62207605	14302848	A	T	1.092	7.79×10^{-2}	MACROD2
	20:14325917_GTTATTA_G	14325917	G	GTTATTA	1.413	2.33×10^{-2}	MACROD2
	rs143540982	14089726	G	A	1.236	1.78×10^{-2}	MACROD2

Table S1: Corresponding SNPs in verified super variants with DRAG. The chr i_j represents the super variant in the *j*th set of the *i*th chromosome; SNP, the rsID of the variant, where available; Prior, known from prior publications addressing common genetic variation linked to COVID-19; OR: odds ratio.

Reference

1. Shelton, J.F., et al., *Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity.* Nature Genetics, 2021. **53**(6): p. 801-808.
2. Li, Z., et al., *Gene expression profiling in autoimmune noninfectious uveitis disease.* The Journal of Immunology, 2008. **181**(7): p. 5147-5157.