Impact of Clinical and Genomic Factors on SARS-CoV2 Disease Severity

Sanjoy Dey, Ph.D.[1], Aritra Bose, Ph.D.[2], Prithwish Chakraborty, Ph.D.[1], Mohamed Ghalwash, Ph.D.[1], Aldo Guzman Saenz, Ph.D.[2], Filippo Utro, Ph.D.[2], Kenney Ng, Ph.D.[1], Jianying Hu, Ph.D.[1], Laxmi Parida, Ph.D.[2], Daby Sow, Ph.D.[1]
[1]Center for Computational Health, IBM Research, Yorktown Heights, NY, USA;
[2]Computational Genomics, IBM Research, Yorktown Heights, NY, USA

Abstract *The SARS-CoV2 virus behind the COVID-19 pandemic is manifesting itself in different ways among infected people. While many are experiencing mild flue-like symptoms or are even remaining asymptomatic after infection, the virus has also led to serious complications, overloading ICUs while claiming more than 2.6 million lives world-wide. In this work, we apply AI methods to better understand factors that drive the severity of the disease. From the UK BioBank dataset we analyzed both clinical and genomic data of patients infected by this virus. Leveraging positive-unlabeled machine learning algorithms coupled with RubricOE, a state-of-the-art genomic analysis framework for genomic feature extraction, we propose severity prediction algorithms with high F_1 score. Furthermore, we extracted insights on clinical and genomic factors driving the severity prediction. We also report on how these factors have evolved during the pandemic w.r.t. significant events such as the emergence of the B.1.1.7 SARS-CoV2 virus strain.*

Introduction

COVID-19, caused by the deadly "severe acute respiratory syndrome coronavirus 2" (SARS-CoV2) virus, is one of the worst pandemic in human history inflicting severe casualties across the globe. So far, it has caused 117,332,262 confirmed cases while claiming 2,605,356 deaths worldwide[1]. This virus has stressed many health systems beyond their limits, screaming for an urgent reaction from research to help mitigate this unprecedented societal threat.

The outbreak of the virus first occurred in Wuhan province of China and then rapidly spread world-wide using human-to-human transmission[2]. One of the primary reason for the rapid spread is due to the wide spectrum of symptoms inhibited by the virus ranging from asymptomatic to mild to rapid progression to critical stage with pneumonia and likely death of respiratory failure[3]. In general, patients with mild symptoms cure easily whereas severe COVID-19 patients require further treatments such as ventilation in ICU. Such heterogeneity in terms of COVID-19 disease pose great challenges for designing treatment protocols. Typically, severe effects of infectious diseases like COVID-19 are hypothesized to be associated with the variations of host genome[4]. Several studies[5] were conducted recently to find the biological mechanisms of the severity of COVID-19 diseases such as genome wide association studies (GWAS) of severe COVID-19[6]. In addition, clinical factors such as demographic, socio-economic factors, lifestyles, prior conditions may also have impact on the spread, exposure, and severity. So does the viral load of the SARS-CoV2 virus. The availability of rich electronic health records (EHR) provides an opportunity for finding such important clinical factors associated with COVID-19 along with important clinical factors observed in EHRs[3,7,8].

Finding the bio-markers for severe COVID-19 patients requires rigorous study on clinical and genomic datasets. Most of the existing studies that aimed at finding the common risk factors for severe COVID-19 patients focused on either the clinical or the genomics factors, but not both. In this study, we aim to find the common risk factors associated COVID-19 severity using diverse factors from both clinical and genomics data available from a large-scale EHR dataset. Such kind of combined clinico-genomic factors can provide a holistic view of COVID-19 disease mechanisms. Specifically, it can provide the additional effects of genomic factors on the host genome as they relate to common clinical factors.

However, there are few challenges in finding the clinical and genomic factors associated with COVID-19 that can be used as bio-markers. *First*, defining the COVID-19 severity from EHR datasets is often challenging, since it is not collected directly in these datasets. Hence, the severity has to be defined using some clinical knowledge as a surrogate phenotype which in turn may introduce label noise and data collection bias. *Second*, assessing the impact of genomic factors on COVID-19 severity may be confounded by several other clinical factors such as the prior comorbidities of patients and treatment protocols. The effect of such confounding factors has to be addressed carefully when conducting the combined clinico-genomic studies. *Third*, the impact of clinico-genomic factors may also vary depending on the variations of diverse COVID-19 strains that have been observed due to genetic mutations.

Table 1: The distribution of unique patients based on Origin and result as of Jan 21, 2021

	Result=0	Result=1	All
Origin=0	11,989	9,083	21,072
Origin=1	31,800	4,345	36,145
All	43,789	13,428	57,217

In this study, we propose a combined framework for finding the clinical and genomic factors associated with COVID-19 severity using a large COVID-19 dataset from the UK Biobank (UKBB). To address the above mentioned challenges, we first use the COVID-19 related hospitalization as a surrogate outcome for defining severe COVID-19 cases. Second, we use a machine learning technique called positive-unlabeled (PU) learning to address the noise and reporting bias present in COVID-19 severity labels. Moreover, we use a recently proposed genomic analysis framework entitled RubricOE[9] to select the set of genomic factors after adjusting for the common prior comorbid conditions which may act as potential confounders. Finally, we aim to assess how the importance of the extracted bio-markers evolve over the pandemic marked by event such as the emergence of the reportedly more contagious B.1.1.7 COVID-19 strain.

Method

Dataset

We analyzed the large prospective cohort of patients from the UK collected in the UKBB[10] repository. The dataset contains diverse information including demographics, diagnosis, medications, lab tests, and genomic information of approximately half a million patients. In addition, the national SARS-CoV-2 laboratory test data were made available in UKBB through the Public Health England (PHE). This COVID-19 dataset contains a flag indicating *specimen origin* of COVID-19 tests: hospital inpatient origin vs other settings. It also notes the specimen collection date and the *specimen result* in addition to a few other information about how the specimen was collected. Data were available from March 16, 2020 and the repository has been updated once or twice every month. In our current study, we used data from March 16, 2020 to January 21, 2021.

Preparing the outcome and feature sets

Although the COVID-19 dataset in UKBB is a prospective study, it did not collect the COVID-19 severity explicitly. The nature of this dataset is very diverse depending on national testing strategy which evolved over time as the pandemic progressed. For example, UK testing was initially restricted to those with symptoms in hospital and thus under this assumption, using the positive tests of these subjects in hospital can be a reasonable proxy for severity. However, when the testings were expanded covering even asymptomatic patients from diverse facilities, defining the severity required analyzing information regarding the specimen sources of test.

Among several different fields available in the COVID-19 table of our the UKBB, we settled on *specimen origin* and *specimen results* to track severity. We have used a combination of these flags for disease severity under the assumption that patients with a positive COVID-19 test obtained in hospital are likely to be severe cases. Table 1 contains the number of unique patients who had at least one positive test result and at least one 'inpatient' indicator retrieved from the origin field. We used inpatient samples COVID-19 patients (Origin=1 and result =1) as the severe COVID-19 patients while other COVID-19 patients (Origin=1 and result = 0) as the non-severe COVID-19 cases.

While this approach estimates severity, it may be prone to a selection bias problem by defining all inpatient COVID-19 patients as severe cases. For instance, some patients may have been hospitalized due to other conditions or may have been experiencing other problems before they got the COVID-19 infection. To cope with aspects of this problem, a separate field covering hospitalization-acquired infections was collected in the database to filter out such patients from actual severe COVID-19 infection related hospitalization. However, this field was enabled in the dataset very recently and the protocol for marking that field is not yet uniformly applied to all samples.

Another potential issue with the dataset pertains to defining the non-severe COVID-19 patient cohort. While hospitalization data for acute COVID-19 patients were collected carefully, the non-severe patients are harder to define. For

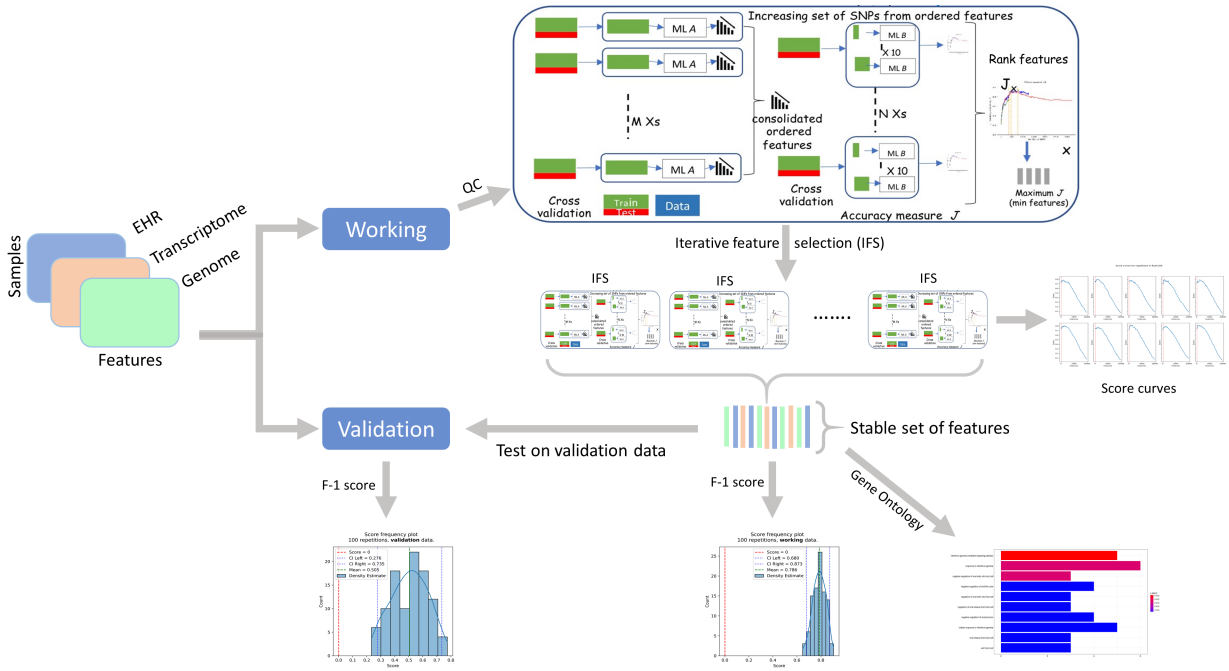


Figure 1: Components of RubricOE, as ensemble ML pipeline extracting stable features from multi-omics data.

example, patients who were tested positively for COVID-19 in outpatients or non-urgent care service centers (Origin=0) may have experienced severe symptoms. The lone presence of the Origin flag set to 0 does not always mean that patients did not have severe reaction at a later stage - they may not have been traced prospectively in the dataset. Such tracing would most likely require conducting another test while being in an inpatient service. To cope with this issue, we used a special class of machine learning to handle this kind of noise associated in the outcome labels.
To build our severity prediction models, we used several clinical and genomic features that may have an impact on COVID-19 severity. For the clinical features, we manually curated some of the most relevant features which may have association with COVID-19 for further analysis. Our clinical features include demographics, lifestyles, comorbidities based on patient's prior history, self-reports, and their chronic disease histories curated from prior inpatient records.

Selecting Genomic Features

Starting with 12,965 overlapping samples with genomic information in the UKBB COVID-19 dataset, we only retained high quality imputed genomic markers in the form of Single Nucleotide Polymorphism (SNPs) with imputation quality INFO score > 0.3, resulting in approximately 4.4 million SNPs. We further did a standard quality control (QC) for filtering informative variants, removing SNPs as well as individuals with more than 2% of missing data. We further filtered for Minor Allele Count (MAC) > 2, Minor Allele Frequency (MAF) > 0.05, sex discrepancy in individuals, variants deviating by more than $p < 1e-6$ in the Hardy-Weinberg equilibrium (HWE) for cases and $p < 1e-10$ in controls. We also filtered for individuals with ± 3 standard deviation in heterozygosity rates and very closely related individuals (second degree). All of the QC analysis was done using PLINK v1.9[11]. We ended up with 12,389 individuals and 4,539,795 SNPs which passed all the QC thresholds. Of these 12,389 individuals, 4,000 were cases with severe COVID-19 infection and 8,389 were controls or non-severe individuals who tested positive. We further pruned for linkage disequilibrium (LD), removing correlated SNPs with $r^2 > 0.2$ with a variance inflation factor of 10 on a window size of 50 kb. We used this pruned dataset for Principal Component Analysis (PCA) and applying downstream algorithms to select stable features.

Genome-wide association study We performed a GWAS on the filtered dataset to perform a sanity check for the QC analysis as well as observe significant loci associated with the severity of COVID-19 infection. GWAS was performed by the package SAIGE[12], which accounts for case-control imbalance in the data. We performed genome-wide association tests while correcting for known confounders in genomics such as population structure, sex and age. We computed the top 20 principal components (PCs) with TeraPCA[13] as a proxy for population structure and used them as covariates along with the biomarkers[14].

Stable feature selection We applied RubricOE (learning rubric for multi-omics and genetic epidemiology), a framework for learning stable features discriminating cases from controls to understand significant SNPs associate with COVID-19 severity. The rubric employs a nested test-training set configuration on the genomic data after QC and pruning for LD. The outer test-training set split reserves the test set ("validation") for final SNP evaluation and denotes the training set as "working" data. Within the latter, further train-test splits are performed to rank the features by their Youden index[15] (also known as J statistic). The pipeline of RubricOE is outlined in Figure 1. RubricOE is applied on the genomic data, accounting for top 20 PCs, sex, year of birth and Diabetes Mellitus (DM) as covariates. It iteratively finds a "stable" set of SNPs which persistently remains highly ranked with their respective Youden index.

Functional annotation

Annotation, Gene Ontology (GO) and enriched pathway analysis of the "stable" set of SNPs produced by RubricOE has been evaluated using the R package clusterProfiler[16] v3.12.

Framework

We used a predictive learning framework called Positive-unlabeled learning (PU learning) for extracting the most significant clinical and genomic features associated with COVID-19 outcome. Note that our dataset is more robust in terms of defining the severe COVID-19 patients than the non-severe cases. Hence, using a classical binary classification algorithm will not be appropriate and able to cope with such noisy class labels. PU learning algorithms have alternatively been reported to be especially suitable for learning from a noisy negative class by treating instances in this class as unlabeled while predicting the positive class (severe COVID-19 cases).

PU learning is a variant of the classical binary classification problem, where it is assumed that the unlabeled examples could belong to either positive or negative class. We refer to[17] for a comprehensive review of such PU learning techniques. Among several state-of-the-art PU learning techniques, we use PU Adapter[18], uPU[19], and nnPU[20] for learning the COVID-19 severity. All of these PU learning techniques have the basic assumption that the observed samples are drawn uniformly from the positive distribution, which enables PU learning techniques to leverage standard binary classification methods with minor modification of data or algorithm.

The PU-Adapter[18] technique pre-processes the available PU data so that any standard binary classification method can be used for learning on the PU dataset. In particular, this method reweighs the samples of PU data so that learning a traditional binary classifier on the weighted samples yield to similar target probability threshold on the PU dataset. Moreover, it has been shown that the classifier trained on positive and unlabeled examples predicts probabilities that differ by a constant factor from the true probabilities of positive class. Unbiased PU learning (uPU)[19] uses two separate loss functions for the positive and unlabeled samples using convex optimization framework.

Evaluation

We evaluate the effectiveness of PU-learning algorithms with a comparison with baseline traditional classification algorithm using logistic regression with a L_1 loss. Computing traditional metrics such as precision, recall, F_1 score, and accuracy is challenging in the positive unlabeled scenario since the only information available relates to the positive label while no sample from the negative class is clearly provided. One straightforward solution to this problem is to assume that the unlabeled data are negative. However, this is not fully accurate. A modified F_1 score (we call it as F_1^{pu} score) has been proposed to address this issue[21]. This score is computed as $\frac{r^2}{p(\hat{y}=1)}$, where r is the recall and $p(\hat{y}=1)$ is the prevalence of the predicted positive label. As shown in[21], the F_1^{pu} score can take values greater than one but

shares the same property as F_1 score; it is high when both precision and recall are high and small when either one is small. We also report metrics used to evaluate traditional classification methods for completeness although these are not recommended for our setting. They include accuracy, precision, recall, and F_1 score.

We used a 5-fold cross-validation (CV) approach for evaluating the performances of the predictive models, where in each CV fold, the hyper-parameter for L_1 loss was tuned internally using a grid-search over the range of $[10^{-3}, 10^{-2}, 10^{-1}, 1, 10]$. Finally, we used a bootstrapping technique for estimating the confidence intervals of the scores using 100 bootstrapping runs for finding the importance of the clinico-genomic features.

Results

The results of the baseline L_1 regularized logistic regression are shown in Table 2. We report the metrics for three different models: clinical, genomic, and the combined clinico-genomic model. Overall, we can observe that clinical model has worse predictive power than the model built from genomic features obtained from GWA data. However, the combined model has significant improvement in terms of predictive power.

Table 3 and Table 4 show the same metrics from the three different models using two different PU learning techniques: nnPU and PU-Adapted logistic regression, respectively. The nnPU model does not perform well for categorical genomic features, moreover it is harder to interpret such model. In contrast, the PU-adapted method uses logistic regression as the baseline model. Hence it is easier to interpret the resulting coefficients as log-odds ratios of clinico-genomic features. The combined clinico-genomic model of the PU adapted logistic regression model yields significant improvement from the models built on clinical and genomic model individually. Although the F_1 score of the PU learning model is similar to the score of the baseline logistic regression model, the adjusted F_1 score is higher for the PU model, which justifies the use of the PU model for our noisy class level.

We interpret the log-odds of the PU-Adapted logistic regression model to denote the importance of a particular feature toward COVID-19 severity. Figure 2 shows the feature importance of top 50 statistically significant features associated with COVID-19 severity computed by the final combined clinico-genomic model using the PU learning framework. The statistical significance of a feature is computed using bootstrapping with 100 runs and the confidence interval is represented as error bars in the figure which are beyond zero (the vertical line).

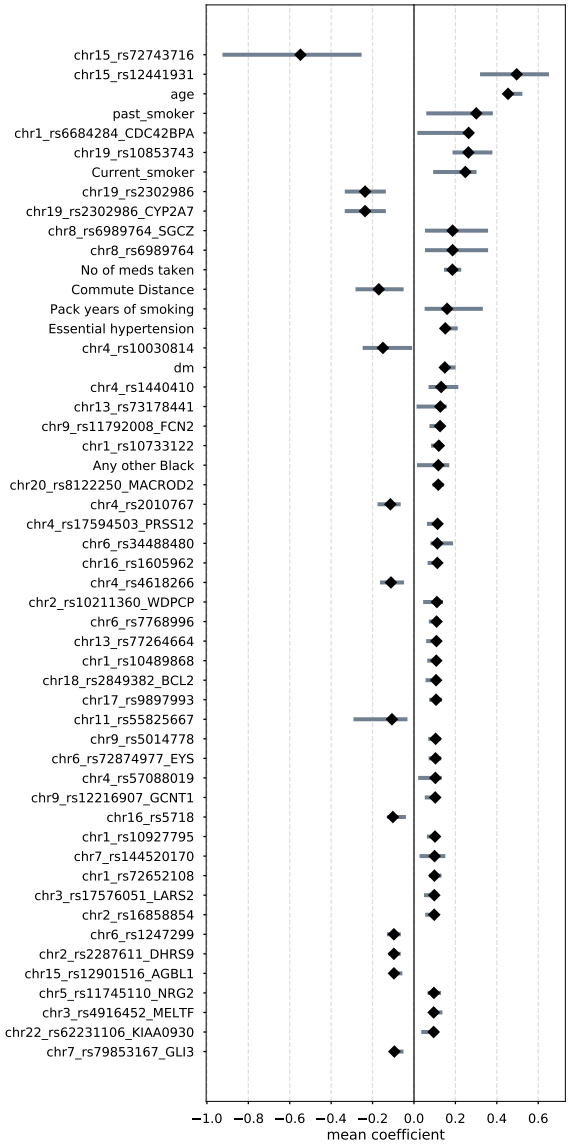


Figure 2: Top statistically significant features obtained from PUAdapted Logistic Regression. Plots show mean coefficients (log odds ratio) for regression along with 95% bootstrapped confidence interval.

Discussion

Our final combined clinico-genomic model finds the significant features associated with COVID-19 severity (Figure 2), which may act as potential biomarkers. Some of the top significant clinical features have already been found in previous studies. For example, age and black ethnicity have been reported in[8], and prior conditions like diabetes mellitus, hypertension have been reported in[7,22,23]. Note that our study finds smoking status (past smoker, current

Table 2: Average (standard deviation) 5-fold CV performances of three models based on Logistic Regression with L_1 loss

	Precision	**Recall**	**Accuracy PU**	F_1 **score**	**Accuracy**	F_1^{pu} **score**
Clinical	0.488 (0.008)	0.629 (0.014)	0.768 (0.006)	0.550 (0.005)	0.667 (0.005)	0.949 (0.019)
Genomic	0.507 (0.023)	0.665 (0.013)	0.799 (0.018)	0.575 (0.019)	0.684 (0.009)	1.045 (0.035)
Clinico-Genomic	0.564 (0.023)	0.710 (0.030)	0.875 (0.022)	0.629 (0.026)	0.730 (0.009)	1.241 (0.057)

Table 3: Average (standard deviation) 5-fold CV performances of three models based on nnPU Model

	Precision	**Recall**	**Accuracy PU**	F_1 **score**	**Accuracy**	F_1^{pu} **score**
Clinical	0.586 (0.017)	0.416 (0.019)	0.803 (0.003)	0.486 (0.009)	0.716 (0.006)	0.754 (0.022)
Genomic	0.4 (0.545)	0 (0.000)	0.677 (0)	0.001 (0.001)	0.677 (0.000)	0.004 (0)
Clinico-Genomic	0.830 (0.025)	0.054 (0.010)	0.694 (0.003)	0.100 (0.018)	0.691 (0.003)	0.138 (0.027)

Table 4: Average (standard deviation) 5-fold CV performances of three models based on PU Adapted Logistic regression Learning

	Precision	**Recall**	**Accuracy PU**	F_1 **score**	**Accuracy**	F_1^{pu} **score**
Clinical	0.401 (0.008)	0.800 (0.011)	0.621 (0.022)	0.534 (0.006)	0.548 (0.015)	0.991 (0.028)
Genomic	0.417 (0.004)	0.862 (0.006)	0.660 (0.010)	0.562 (0.005)	0.567 (0.007)	1.114 (0.018)
Clinico-Genomic	0.485 (0.006)	0.839 (0.021)	0.794 (0.012)	0.615 (0.009)	0.660 (0.007)	1.260 (0.040)

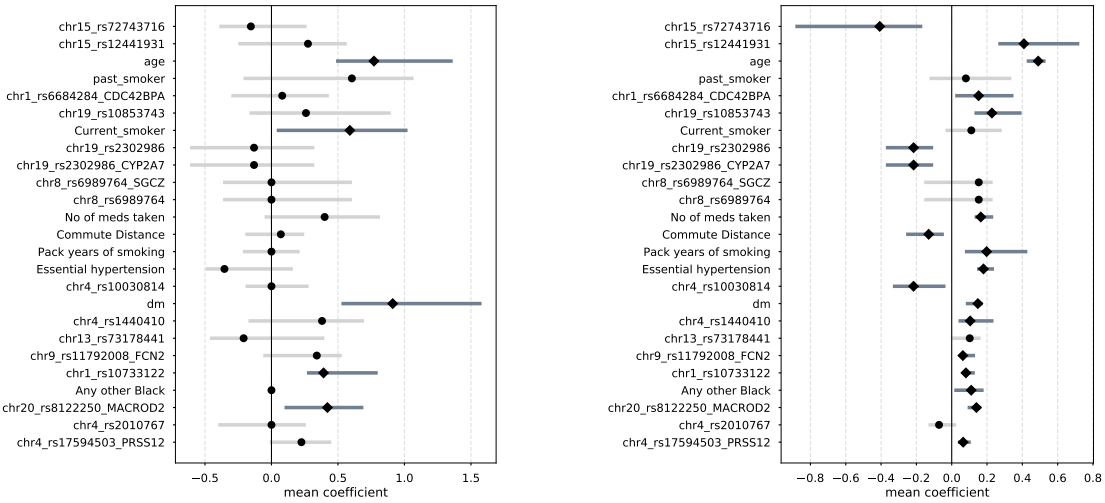
smoker, pack years of smoking) is related to severe COVID-19 patients, which has ambiguous evidence in literature. A few studies[8] reported it with increased risk factors, while others[22] did not find it to be a significant risk factors for COVID-19 severity.

In addition, our study found a few more features such as number of medication taken per day to be significant. Such features are related to the overall health status of the patient and correlate indirectly to high severity risk factors. Interestingly, distance from home to workplace has also been reported with lower risk of severity, which may be confounded by people commuting by car for longer distance instead of public transport or cycling for shorter commuting distance.

The genomic features obtained from RubricOE are enriched in a host of biological pathways related to transmembrane transporters and receptor activities (Figure 5(a)). The most significant is *transmembrane receptor protein tyrosine phosphatase* which on increasing activity might affect T cells and contribute to their depletion and immunoparalysis in severe COVID-19 patients[24]. Other important pathways such as *ion channel*, *metal ion* and *inorganic cation* transmembrane transporter activities reaffirm the notion that SARS-CoV-2 E protein is a potential ion channel[25] like other coronaviridae[26]. Focusing on the significant biomarkers from the combined model, we observed two new GO enriched terms, *passive transmembrane transporter activity* and *PDZ domain binding* (Figure 5(b)). It has been shown that PDZ-containing proteins among binders of the SARS-CoV-2 proteins E, 3a or N affect viral replication under knock-down gene expression in infected cells, significantly[27], particularly in the E protein[28].

The impact of clinico-genomic features can be confounded by different treatment protocols. However, our UKBB dataset does not have direct information on COVID-19 drugs and detailed treatment protocols that a COVID-19 patient went through, so we tried to assess the impact of a COVID-19 drug indirectly using its date of approval. In particular, we assessed how the co-efficients of topmost features mentioned in Figure 2 change before and after the drug was approved for treatment. As a simple use-case, we report the changes of impact of the top 25 biomarkers (instead of all 50 due to space limitation) for the first approved UK drug called 'Dexamethazone' with its approval date as June 15, 2020 in Figure 3. We can observe that most of the 25 significant factors from the whole dataset still had similar impact after the drug being used, but only 5 features had significant impact before the drug was used.

Similarly, we also wanted to assess how the impacts of clinico-genomic features change over different strains of SARS-CoV-2. In the UK, a reportedly more contagious strain named B.1.1.7[29] started surfacing from early September 2020 and it is estimated that by December 2020, over 60% of new COVID-19 patients had the newer strain. The UKBB also



(a) before approval (N=1,756) **(b)** after approval (N=10,633)

Figure 3: Mean coefficients with 95% CI of features from Fig 2 indicating severity of COVID-19 (a) before and (b) after Dexamethasone was approved for usage in UK. Features that are still significant (not significant) are shown via deep grey (light grey) bars with square (circular) markers.

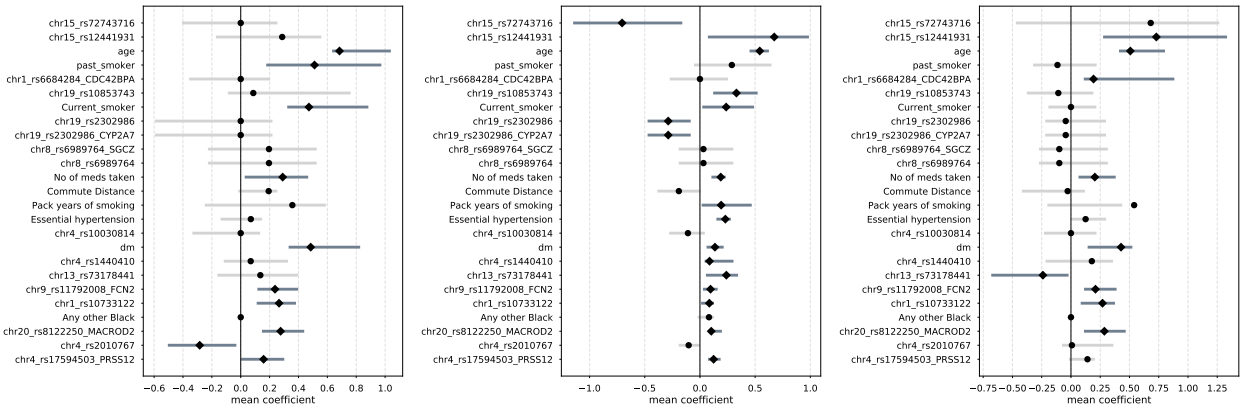
did not collect the virus genome prospectively. So the only way to assess the impact of clinico-genomic biomarkers is through the date when a patient was diagnosed with COVID-19. We divided the whole cohort into three parts: those diagnosed before Aug 31, 2020 (older strain), diagnosed between September 1st, 2020 and December 15, 2020 (both strains) and after December 15, 2020 (newer strain). Figure 4 shows how the effect of the top clinico-genomic features obtained by the global model change over these three cohorts. We can observe that some factors like age, prior DM condition and number of medication are consistent across both strains, while the smoking history (current and present) had larger impact on older strain than the newer strain. Such analysis of temporal trend of a factor's impact on severity can help elucidate important clinical insights.

Related Works

To combat the global pandemic caused by COVID-19, scientists across different disciplines have been studying the disease to understand various facets such as its spread, epidemiological and patho-physiological characteristics, and societal impact. In this section we survey some of the most relevant literature.

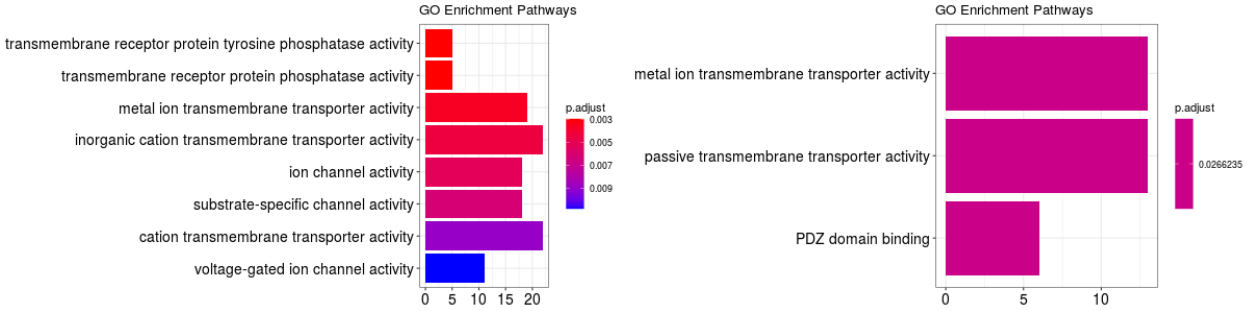
Some of the more studied aspects of understanding the spread and clinical manifestations of the disease has been focused on the viral and host genomic factors[30]. To help consolidate and streamline the efforts around studying the host genetics for susceptibility the COVID-19 Host Initiative was launched in mid 2020[4] with a primary focus on disease severity. As part of this initiative, Hou et al.[31] found that ACE2 or TMPRSS2 DNA polymorphisms were likely associated with genetic susceptibility of COVID-19, while Li et al.[32] found that genetic variants under a polygenic model shows promising improvements in prediction accuracies and argued for greater investigation into polygenic risk scores. Wang et al.[33] presented further genetic insights into phenotypic difference among the COVID-19 patient groups, highlighting genes and variants of interest. More superficially, they conducted both single-variant and gene-based association with respect to five severity groups finding insights such as variants involved in IL-1 signaling pathways to be of interest. While studies identified a gene cluster on chromosome 3 as a risk locus for respiratory failure after infection with COVID-19, Zeberg et al.[34] found that the risk is conferred by a genomic segment that is carried by around 50% of people in South Asia and 16% of people in Europe.

Simultaneously, there has been concerted efforts to characterize the clinical characteristics of COVID-19 including the patho-physiology and the risk factors for covid. Several large-scale meta-analysis studies were conducted either on previous studies (Fu et al.[3]) or by summarizing published articles (e.g., Li et al.[7], Zheng et al.[8]) to identify com-



(a) older strain **(b)** both strains **(c)** newer strain

Figure 4: Mean coefficients with 95% CI of features from Fig 2 indicating severity of COVID-19 with dominance of (a) older strain, (b) both strains, and (c) newer strain (B.1.1.7) in UK. Features that are still significant (not significant) are shown via deep grey (light grey) bars with square (circular) markers.



(a) Pathways of "stable" SNPs from RubricOE. **(b)** Pathways of significant SNPs from combined model.

Figure 5: Boxplots of significant GO enriched pathways, colored by their corresponding corrected p-value and count of mapped genes in the x-axis.

mon clinical symptoms and lab abnormalities among severe COVID-19 patients. Some common risk factors observed among critical/mortal COVID-19 patients are socio-demographic information such as male, older than 65, smoking, and prior history of hypertension, diabetes, cardiovascular disease, respiratory diseases, kidney diseases, obesity, immunosuppresion, and cancer. In addition to these prior conditions, depression as well as other cognitive and neurological disorders were also found as risk factors by a nationwide cohort study in Israel [22], although this study found that smoking and presence of respiratory diseases do not significantly increase the risk of complications. Williamson[23] presented a risk analysis using a secure health analytics platform (OpenSAFELY) that covers 40% of all patients in England. They found risk factors such as male, age, socio-economic deprivation, diabetes, severe asthma, and various other medical conditions. They also found Black and South Asian people to be at higher risk. Research has also been conducted to characterize risk among sub-populations such as obesity, heart failure, and Parkinson's Disease[35–37] concurrently. Mathew et al.[38] found three immunotypes revealing different patterns of lymphocyte responses among hospitalized COVID-19 patients.

Besides finding such clinico-genomic risk factors of severe COVID-19, several machine learning (ML) and artificial intelligence (AI) have also been applied successfully to solve a wide range of needs, ranging from fast ML based COVID-19 infection prediction [39] to routine blood-tests[40] as fast alternatives to costly and time-consuming PCR tests[41] to predicting disease state of individual cells given their transcriptomes[42].

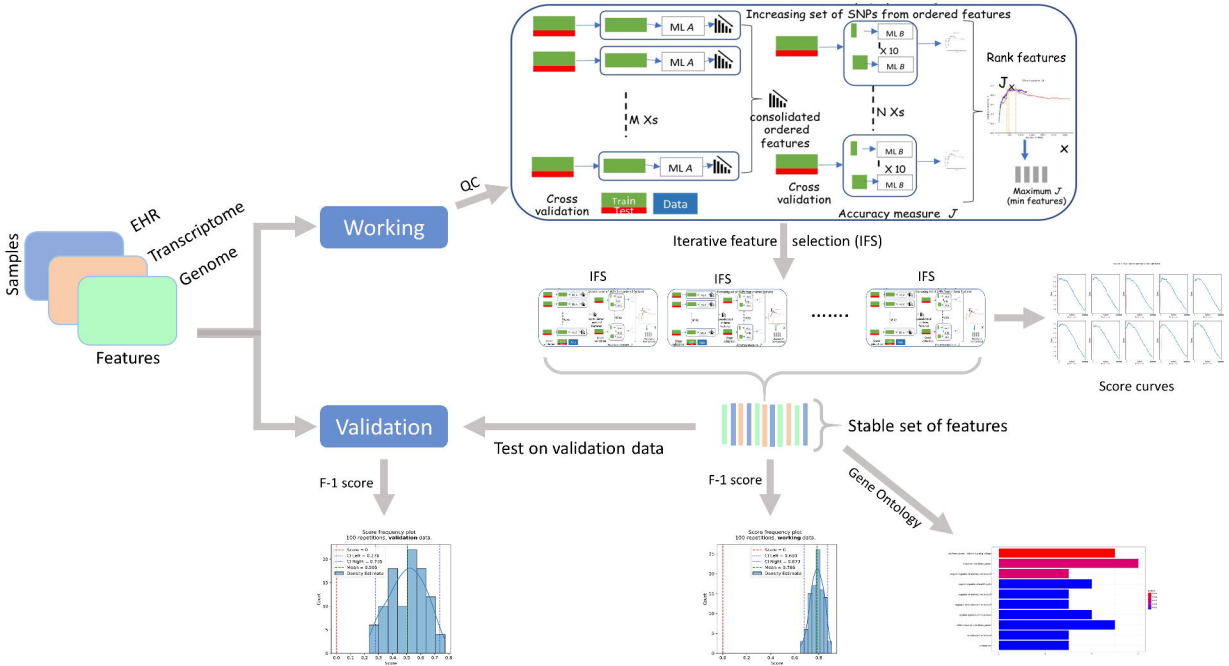
Conclusion

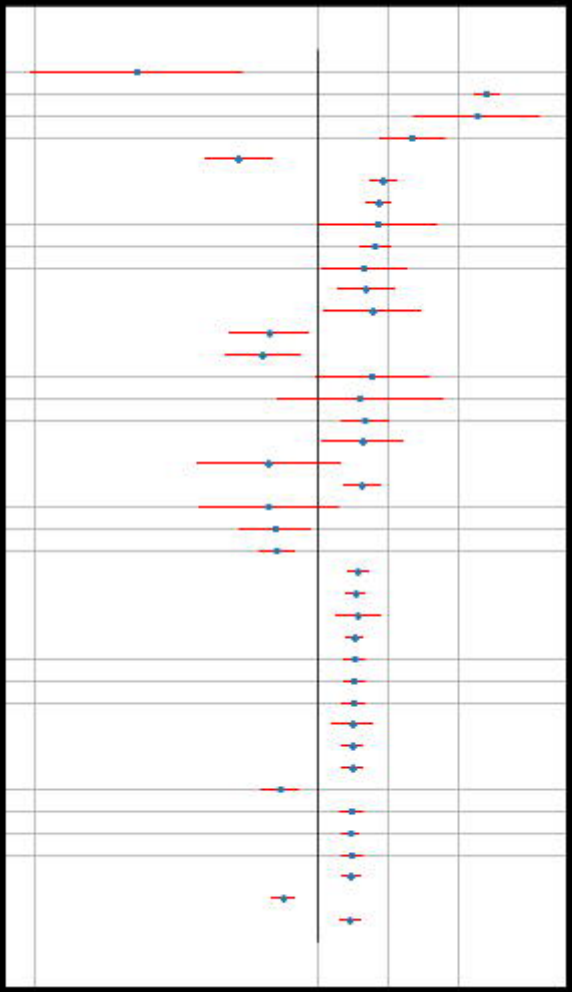
In this study, we looked for significant clinical and genomic factors that are associated with severe reaction of SARS-CoV-2 virus infection from a large-scale EHR and GWA dataset available from the UK. In particular, we used a special class of machine learning called positive-unlabeled learning to address the challenge of noisy class label of COVID-19 severity outcome. Our holistic clinico-genomic modeling can potentially discover the effect of such factors in a robust manner, which can be used as potential biomarkers for better clinical decision making. In future, we plan to further investigate the effect of treatment protocols and vaccination on COVID-19 severity leveraging the inpatient data. Also, investigating risk factors associated with COVID-19 induced death will be an important future research.

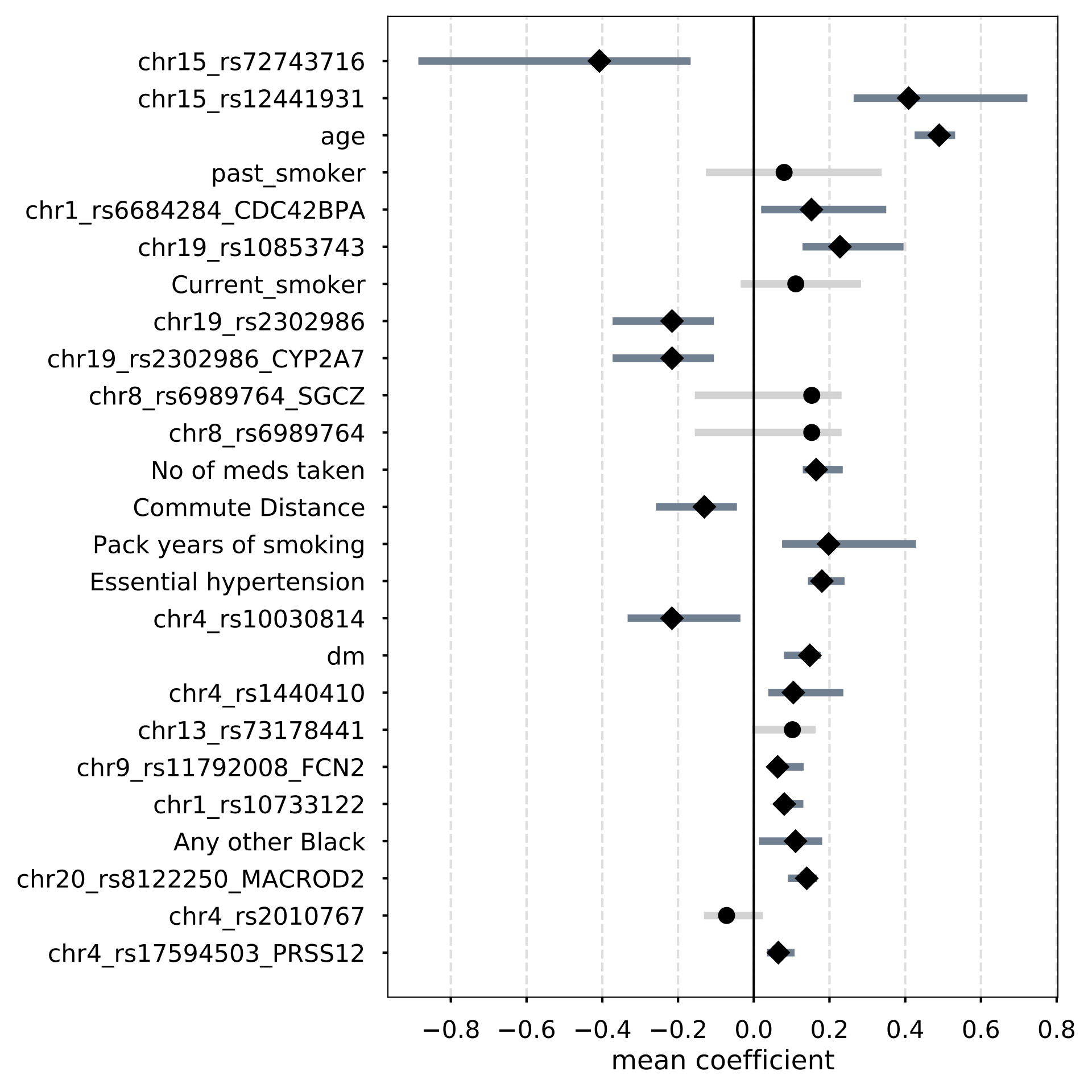
References

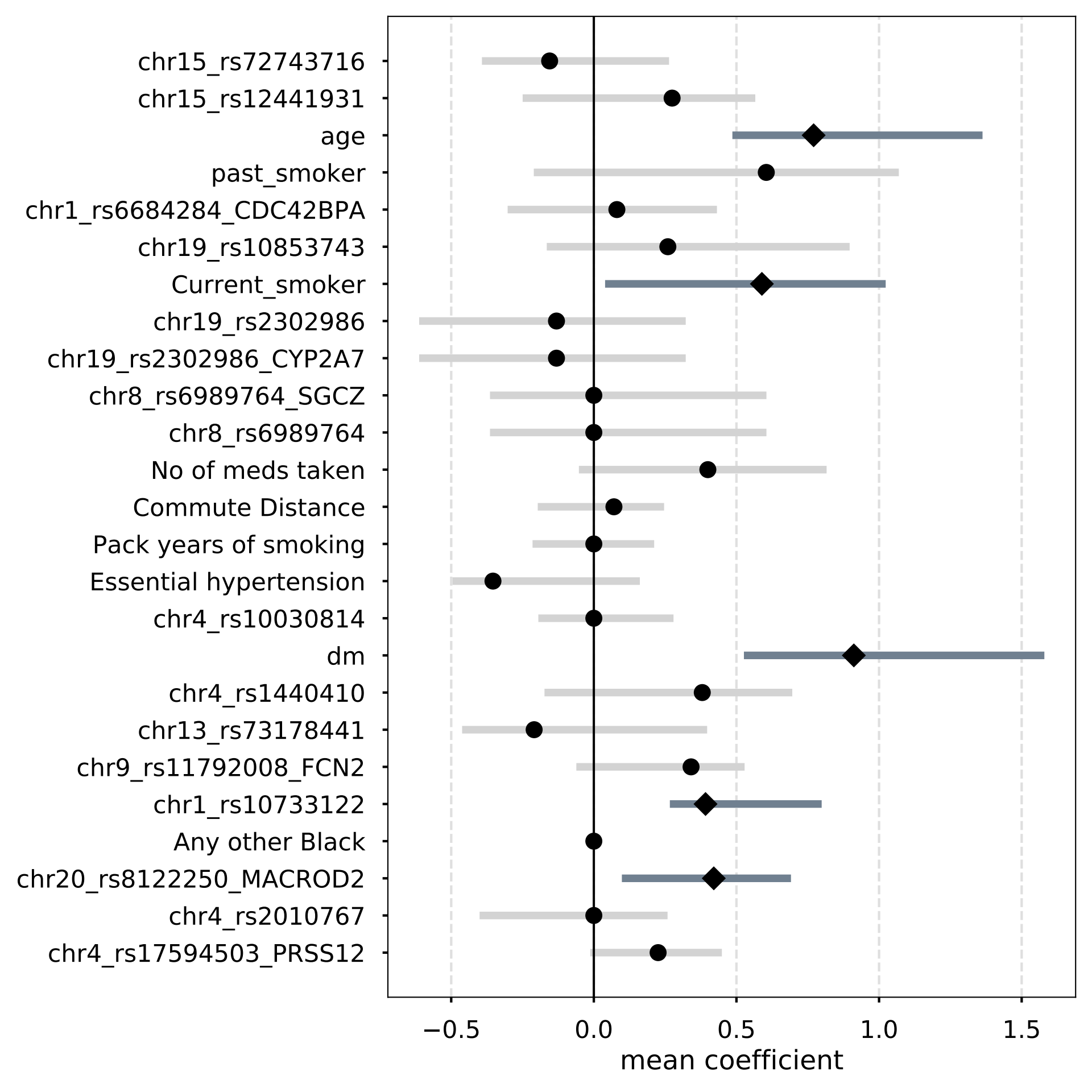
1. WHO;. Last accessed: 2021-03-10. https://covid19.who.int/.
2. Jiang S, Du L, Shi Z. An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies. Emerging microbes & infections. 2020;9(1):275–277.
3. Fu L, Wang B, Yuan T, Chen X, Ao Y, Fitzpatrick T, et al. Clinical characteristics of coronavirus disease 2019 (COVID-19) in China: a systematic review and meta-analysis. Journal of Infection. 2020;80(6):656–665.
4. Initiative CHG, et al. 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):715.
5. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in Covid-19. Nature. 2021;591(7848):92–98.
6. Group SCG. Genomewide association study of severe Covid-19 with respiratory failure. New England Journal of Medicine. 2020;383(16):1522–1534.
7. Li J, Huang DQ, Zou B, Yang H, Hui WZ, Rui F, et al. Epidemiology of COVID-19: A systematic review and meta-analysis of clinical characteristics, risk factors, and outcomes. Journal of medical virology. 2021;93(3):1449–1458.
8. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. Journal of Infection. 2020.
9. Saha S, Guzman-Saenz A, Bose A, Utro F, Platt DE, Parida L. RubricOE: a learning framework for genetic epidemiology. medRxiv. 2021.
10. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 med. 2015;12(3):e1001779.
11. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742–015.
12. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics. 2018;50(9):1335–1341.
13. Bose A, Kalantzis V, Kontopoulou EM, Elkady M, Paschou P, Drineas P. TeraPCA: a fast and scalable software package to study genetic variation in tera-scale genotypes. Bioinformatics. 2019;35(19):3679–3683.
14. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904–909.
15. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.
16. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology. 2012;16(5):284–287.
17. Bekker J, Davis J. Learning from positive and unlabeled data: A survey. Machine Learning. 2020;109(4):719–760.
18. Elkan C, Noto K. Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining; 2008. p. 213–220.
19. Du Plessis M, Niu G, Sugiyama M. Convex formulation for learning from positive and unlabeled data. In: International conference on machine learning. PMLR; 2015. p. 1386–1394.

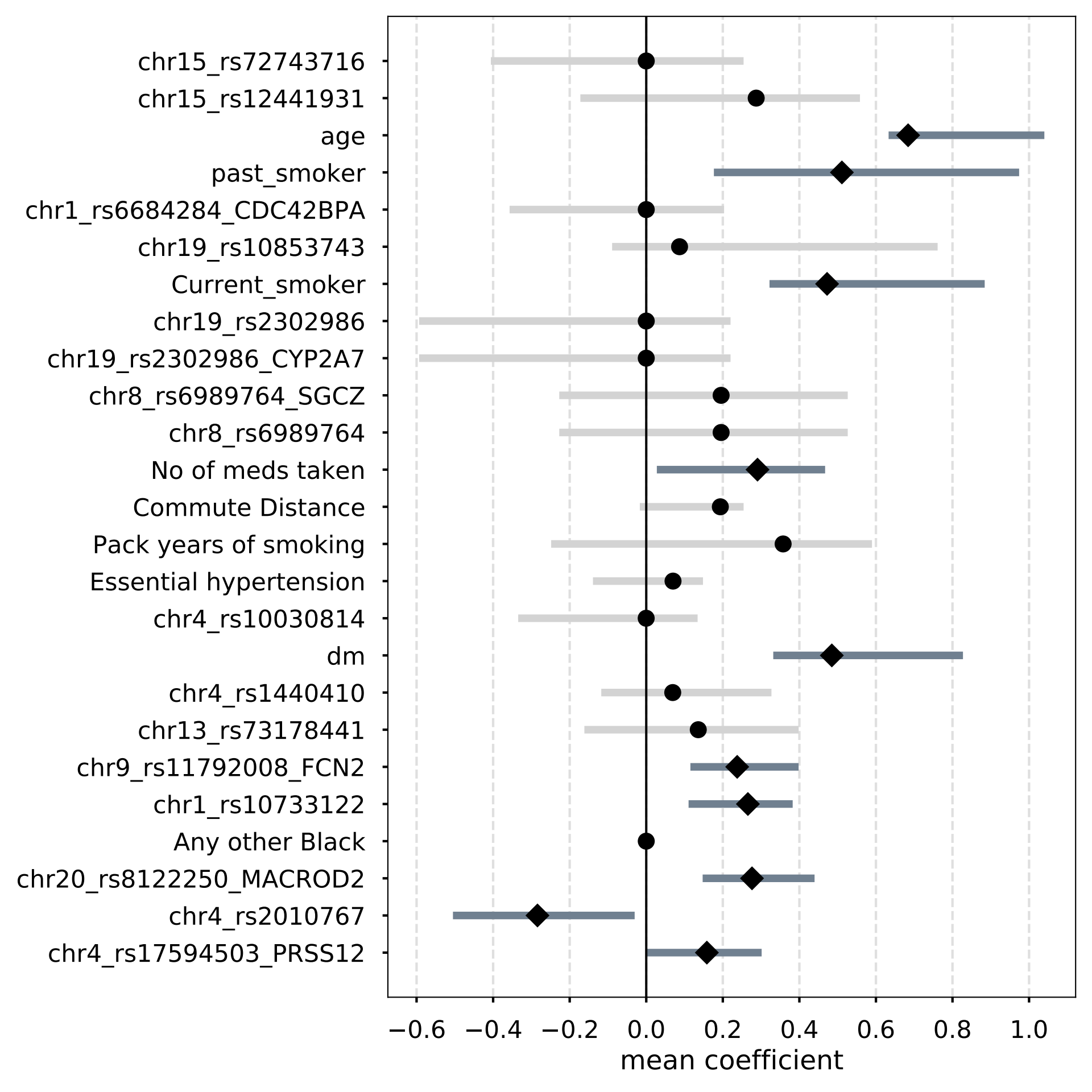
20. Kiryo R, Niu G, du Plessis MC, Sugiyama M. Positive-unlabeled learning with non-negative risk estimator. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. p. 1674–1684.
21. Lee WS, Liu B. Learning with positive and unlabeled examples using weighted logistic regression. In: ICML. vol. 3; 2003. p. 448–455.
22. Yanover C, Mizrahi B, Kalkstein N, Marcus K, Akiva P, Barer Y, et al. What Factors Increase the Risk of Complications in SARS-CoV-2–Infected Patients? A Cohort Study in a Nationwide Israeli Health Organization. JMIR Public Health and Surveillance. 2020;6(3):e20872.
23. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584(7821):430–436.
24. Diao B, Wang C, Tan Y, Chen X, Liu Y, Ning L, et al. Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). Frontiers in immunology. 2020;11:827.
25. Tomar PPS, Arkin IT. SARS-CoV-2 E protein is a potential ion channel that can be inhibited by Gliclazide and Memantine. Biochemical and Biophysical Research Communications. 2020;530(1):10–14.
26. Nieto-Torres JL, DeDiego ML, Verdiá-Báguena C, Jimenez-Guardeño JM, Regla-Nava JA, Fernandez-Delgado R, et al. Severe acute respiratory syndrome coronavirus envelope protein ion channel activity promotes virus fitness and pathogenesis. PLoS Pathog. 2014;10(5):e1004077.
27. Caillet-Saguy C, Durbesson F, Rezelj VV, Gogl G, Tran QD, Twizere JC, et al. Host PDZ-containing proteins targeted by SARS-Cov-2. bioRxiv. 2021.
28. De Maio F, Lo Cascio E, Babini G, Sali M, Della Longa S, Tilocca B, et al. Microbes and Infection. 2020;22(10):592–597. Available from: https://www.sciencedirect.com/science/article/pii/S1286457920301532.
29. B.1.1.7;. Last accessed: 2021-03-09. https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/a-new-strain-of-coronavirus-what-you-should-know.
30. Nexstrain;. Last accessed: 2021-03-09. https://nextstrain.org/sars-cov-2/.
31. Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, et al. New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. BMC medicine. 2020;18(1):1–8.
32. Li R, Chen Y, Ritchie MD, Moore JH. Electronic health records and polygenic risk scores for predicting disease risk. Nature Reviews Genetics. 2020:1–10.
33. Wang F, Huang S, Gao R, Zhou Y, Lai C, Li Z, et al. Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. Cell discovery. 2020;6(1):1–16.
34. Zeberg H, Paabo S. The major genetic risk factor for severe COVID-19 is inherited from Neandertals. BioRxiv. 2020.
35. Yonas E, Alwi I, Pranata R, Huang I, Lim MA, Gutierrez EJ, et al. Effect of heart failure on the outcome of COVID-19—A meta analysis and systematic review. The American Journal of Emergency Medicine. 2020.
36. Bhidayasiri R, Virameteekul S, Kim JM, Pal PK, Chung SJ. COVID-19: an early review of its global impact and considerations for Parkinson's disease patient care. Journal of movement disorders. 2020;13(2):105.
37. Brundin P, Nath A, Beckham JD. Is COVID-19 a perfect storm for Parkinson's disease? Trends in Neurosciences. 2020.
38. Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, Wu JE, et al. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science. 2020;369(6508).
39. McElvaney OJ, Hobbs BD, Qiao D, McElvaney OF, Moll M, McEvoy NL, et al. A linear prognostic score based on the ratio of interleukin-6 to interleukin-10 predicts outcomes in COVID-19. EBioMedicine. 2020;61:103026.
40. Yang HS, Hou Y, Vasovic LV, Steel P, Chadburn A, Racine-Brzostek SE, et al. Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. medRxiv. 2020.
41. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. The Lancet infectious diseases. 2020.
42. Sehanobish A, Ravindra NG, van Dijk D. Gaining insight into SARS-CoV-2 infection and COVID-19 severity using self-supervised edge features and Graph Neural Networks. arXiv preprint arXiv:200612971. 2020.

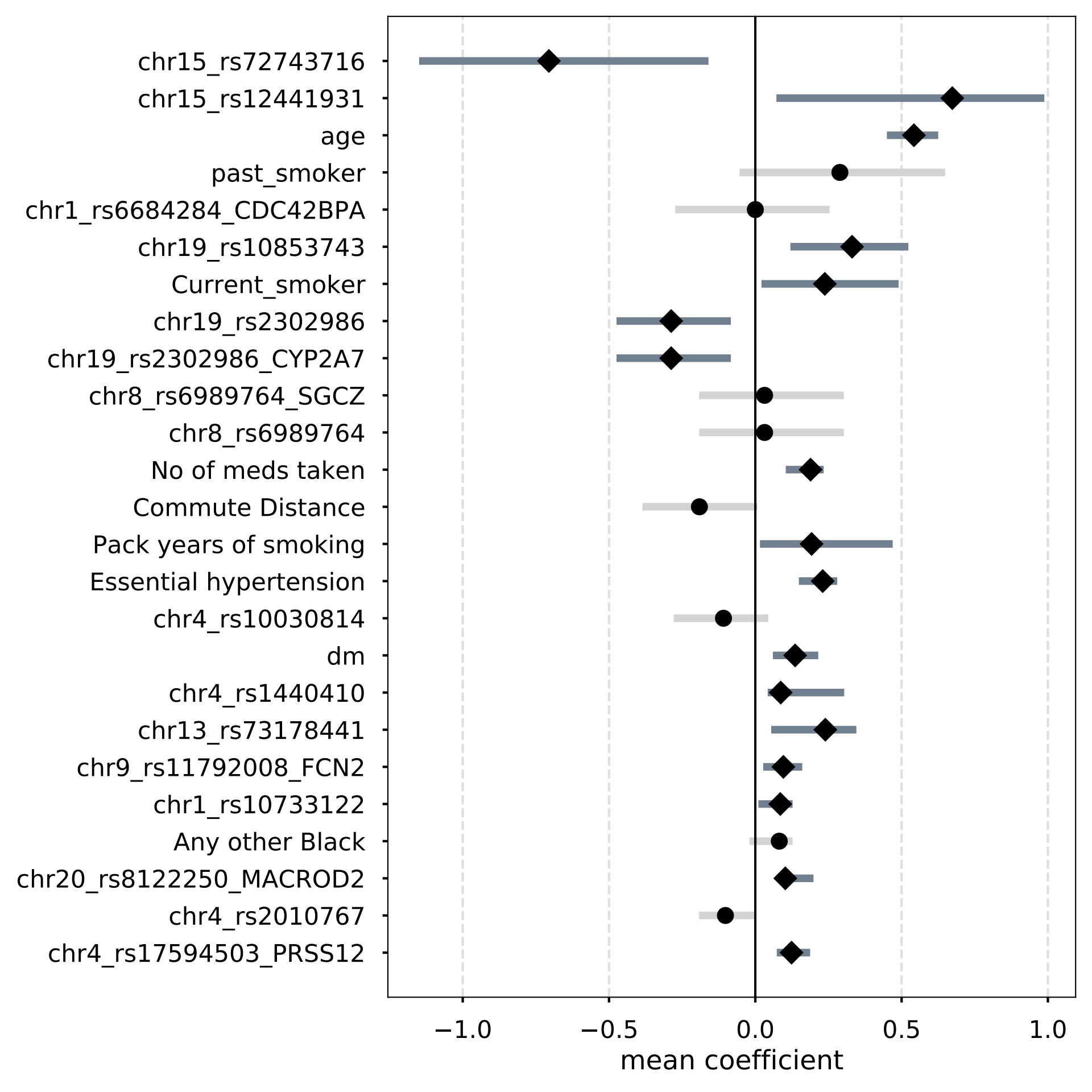

Samples
EHR
Transcriptome
Genome
Features
Working
QC
Increasing set of SNPs from ordered features
ML A
M Xs
consolidated ordered features
ML B
X 10
N Xs
Rank features
J
X
Maximum J (min features)
Cross validation
Train
Test
Data
Accuracy measure J
Iterative feature selection (IFS)
IFS
.....
Score curves
Validation
Test on validation data
Stable set of features
F-1 score
F-1 score
Gene Ontology

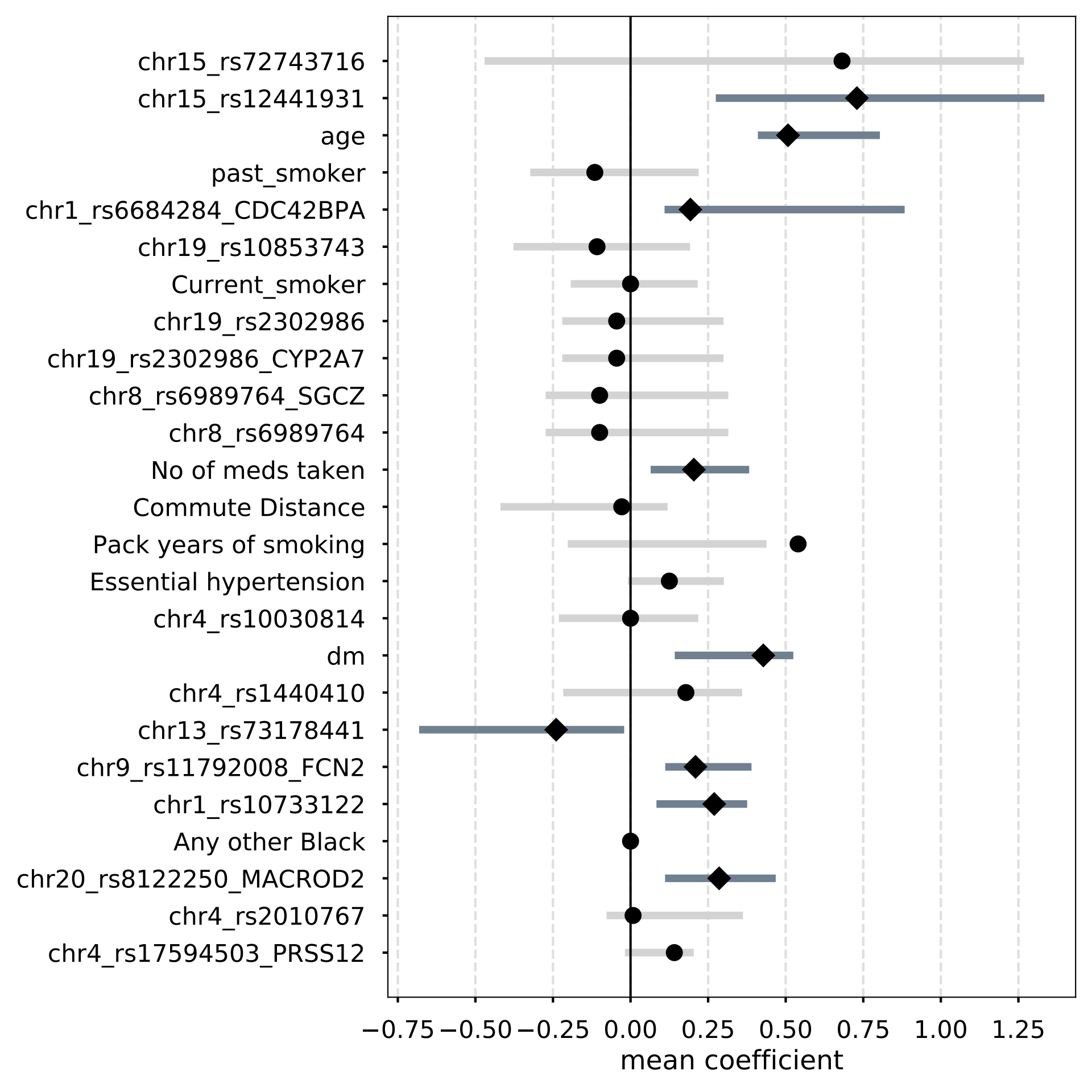


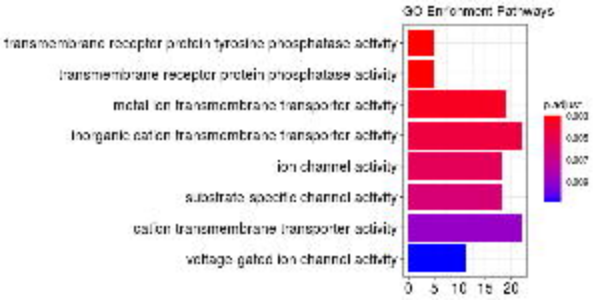

chr15_rs72743716
chr15_rs12441931
age
past_smoker
chr1_rs6684284_CDC42BPA
chr19_rs10853743
Current_smoker
chr19_rs2302986
chr19_rs2302986_CYP2A7
chr8_rs6989764_SGCZ
chr8_rs6989764
No of meds taken
Commute Distance
Pack years of smoking
Essential hypertension
chr4_rs10030814
dm
chr4_rs1440410
chr13_rs73178441
chr9_rs11792008_FCN2
chr1_rs10733122
Any other Black
chr20_rs8122250_MACROD2
chr4_rs2010767
chr4_rs17594503_PRSS12
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
mean coefficient

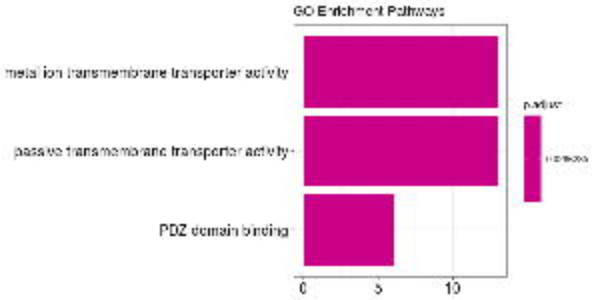

chr15_rs72743716
chr15_rs12441931
age
past_smoker
chr1_rs6684284_CDC42BPA
chr19_rs10853743
Current_smoker
chr19_rs2302986
chr19_rs2302986_CYP2A7
chr8_rs6989764_SGCZ
chr8_rs6989764
No of meds taken
Commute Distance
Pack years of smoking
Essential hypertension
chr4_rs10030814
dm
chr4_rs1440410
chr13_rs73178441
chr9_rs11792008_FCN2
chr1_rs10733122
Any other Black
chr20_rs8122250_MACROD2
chr4_rs2010767
chr4_rs17594503_PRSS12
−0.5
0.0
0.5
1.0
1.5
mean coefficient


chr15_rs72743716
chr15_rs12441931
age
past_smoker
chr1_rs6684284_CDC42BPA
chr19_rs10853743
Current_smoker
chr19_rs2302986
chr19_rs2302986_CYP2A7
chr8_rs6989764_SGCZ
chr8_rs6989764
No of meds taken
Commute Distance
Pack years of smoking
Essential hypertension
chr4_rs10030814
dm
chr4_rs1440410
chr13_rs73178441
chr9_rs11792008_FCN2
chr1_rs10733122
Any other Black
chr20_rs8122250_MACROD2
chr4_rs2010767
chr4_rs17594503_PRSS12
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
mean coefficient


chr15_rs72743716
chr15_rs12441931
age
past_smoker
chr1_rs6684284_CDC42BPA
chr19_rs10853743
Current_smoker
chr19_rs2302986
chr19_rs2302986_CYP2A7
chr8_rs6989764_SGCZ
chr8_rs6989764
No of meds taken
Commute Distance
Pack years of smoking
Essential hypertension
chr4_rs10030814
dm
chr4_rs1440410
chr13_rs73178441
chr9_rs11792008_FCN2
chr1_rs10733122
Any other Black
chr20_rs8122250_MACROD2
chr4_rs2010767
chr4_rs17594503_PRSS12
−1.0
−0.5
0.0
0.5
1.0
mean coefficient


chr15_rs72743716
chr15_rs12441931
age
past_smoker
chr1_rs6684284_CDC42BPA
chr19_rs10853743
Current_smoker
chr19_rs2302986
chr19_rs2302986_CYP2A7
chr8_rs6989764_SGCZ
chr8_rs6989764
No of meds taken
Commute Distance
Pack years of smoking
Essential hypertension
chr4_rs10030814
dm
chr4_rs1440410
chr13_rs73178441
chr9_rs11792008_FCN2
chr1_rs10733122
Any other Black
chr20_rs8122250_MACROD2
chr4_rs2010767
chr4_rs17594503_PRSS12
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
mean coefficient


GO Enrichment Pathways
transmembrane receptor protein tyrosine phosphatase activity
transmembrane receptor protein phosphatase activity
metal ion transmembrane transporter activity
inorganic cation transmembrane transporter activity
ion channel activity
substrate specific channel activity
cation transmembrane transporter activity
voltage gated ion channel activity
0 5 10 15 20
p.adjust
0.009
0.008
0.007
0.006


GO Enrichment Pathways
metal ion transmembrane transporter activity
passive transmembrane transporter activity
PDZ domain binding
0
5
10
p.adjust