Journal of
Translational Medicine

RESEARCH Open Access

Reproducibility of genetic risk factors identified for long COVID using combinatorial analysis across US and UK patient cohorts with diverse ancestries

J. Sardell[1], M. Pearson[1], K. Chocian[1], S. Das[1], K. Taylor[1], M. Strivens[1], R. Gupta[2], A. Rochlin[3†] and S. Gardner[1,3*†]

Abstract

Background Long COVID is a major public health burden causing a diverse array of debilitating symptoms in tens of millions of patients globally. In spite of this overwhelming disease prevalence, staggering cost, severe impact on patients' lives and intense global research efforts, study of the disease has proved challenging due to its complexity. Genome-wide association studies (GWAS) have identified only four loci potentially associated with the disease, although these results did not statistically replicate between studies. A previous combinatorial analysis study identified a total of 73 genes that were highly associated with two long COVID cohorts in the predominantly (> 91%) white European ancestry Sano GOLD population, and we sought to reproduce these findings in the independent and ancestrally more diverse All of Us (AoU) population.

Methods We assessed the reproducibility of the 5343 long COVID disease signatures from the original study in the AoU population. Because the very small population sizes provide very limited power to replicate findings, we initially tested whether we observed a statistically significant enrichment of the Sano GOLD disease signatures that are also positively correlated with long COVID in the AoU cohort after controlling for population substructure.

Results For the Sano GOLD disease signatures that have a case frequency greater than 5% in AoU, we consistently observed a significant enrichment (77–83%, $p < 0.01$) of signatures that are also positively associated with long COVID in the AoU cohort. These encompassed 92% of the genes identified in the original study. At least five of the disease signatures found in Sano GOLD were also shown to be individually significantly associated with increased long COVID prevalence in the AoU population. Rates of signature reproducibility are strongest among self-identified white patients, but we also observe significant enrichment of reproducing disease associations in self-identified black/African-American and Hispanic/Latino cohorts. Signatures associated with 11 out of the 13 drug repurposing candidates identified in the original Sano GOLD study were reproduced in this study.

Conclusion These results demonstrate the reproducibility of long COVID disease signal found by combinatorial analysis, broadly validating the results of the original analysis. They provide compelling evidence for a much broader array of genetic associations with long COVID than previously identified through traditional GWAS studies. This strongly supports the hypothesis that genetic factors play a critical role in determining an individual's susceptibility

[†] A. Rochlin and S. Gardner have joint last authors.

*Correspondence:
S. Gardner
steve@precisionlife.com
Full list of author information is available at the end of the article

to long COVID following recovery from acute SARS-CoV-2 infection. It also lends weight to the drug repurposing candidates identified in the original analysis. Together these results may help to stimulate much needed new precision medicine approaches to more effectively diagnose and treat the disease. This is also the first reproduction of long COVID genetic associations across multiple populations with substantially different ancestry distributions. Given the high reproducibility rate across diverse populations, these findings may have broader clinical application and promote better health equity. We hope that this will provide confidence to explore some of these mechanisms and drug targets and help advance research into novel ways to diagnose the disease and accelerate the discovery and selection of better therapeutic options, both in the form of newly discovered drugs and/or the immediate prioritization of coordinated investigations into the efficacy of repurposed drug candidates.

Keywords Long COVID, Post COVID-19 condition, Post-acute COVID-19, SARS-CoV-2, PASC, Genetics, Reproducibility, Combinatorial analytics

Introduction

Post COVID condition, commonly known as long COVID or PASC (post-acute sequelae of SARS-CoV-2 infection), is a debilitating chronic condition that develops following a SARS-CoV-2 infection in around 5–15% of patients [1]. The global prevalence of long COVID is estimated to be at least 65 million people [2] and is increasing annually. It's estimated to have now affected over 400 million individuals and to cost over $1 trillion (or 1% of global GDP) annually[1], which has caused a long-lasting and profound impact on patients' lives and healthcare systems and has created a major public health issue [3].

'Long COVID' is a term originally defined by patients to describe the post-acute and long-term health effects of COVID-19 [4, 5] and the highly variable symptoms associated with the condition. The frequency and severity of SARS-CoV-2 infections appears to be correlated with increased risk of developing long COVID [6].

Long COVID patients have reported a diverse array of symptoms across multiple organ systems [7] with the most common being post-exertional malaise [8], dysautonomia [9], cognitive dysfunction [10], mood disturbances [11] and respiratory problems [12]. Many of these symptoms and signs are also observed in other complex neuroimmune disorders such as myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) [13–15], postural orthostatic tachycardia syndrome (POTS) [16, 17] and fibromyalgia [18, 19], all of which, like long COVID, disproportionately affect women [20]. To advance our understanding of the pathophysiological mechanisms underlying these shared clinical manifestations, it is important to have a deeper understanding of the genetic similarities between long COVID and other neuroimmune conditions. This effort is hampered, as most of these diseases, like long COVID, are highly complex and have been intractable for existing genomic analysis approaches.

More than four years following the global COVID-19 outbreak, patients still often struggle to obtain a long COVID diagnosis as agreement on the definition of the disease remains elusive beyond self-reported persistence of a wide range of symptoms. Governments also find evaluating its prevalence and setting public health policy difficult due to absence of a clear and consistent definition of the disease [21, 22]. There are currently no recognized laboratory diagnostic tests or disease modifying therapies for long COVID. Research into the biological mechanisms of the disease is hindered by the variability in study designs, lack of reproducible findings across patient populations, and challenges in accurately capturing the heterogenous clinical phenotypes of patient cohorts [23]. A definitive biological explanation of some of the factors causing and defining the disease and a test encompassing these is urgently required to overcome this.

Only a few preliminary GWAS for long COVID have been published to date [24–26], likely due to the challenges of assembling a sufficiently powered patient cohort and the studies' consequently limited findings. A study by the COVID-19 Host Genetics Initiative (HGI) identified only a single significant locus (*FOXP4*) from an analysis of 6450 long COVID cases and over 1 million population controls aggregated from multiple cohorts[25]. Another recent meta-analysis of over 53,000 cases and 120,000 controls from 23andMe identified three significant loci (*HLA-DQA1–HLA-DQB*, *ABO* and *BPTF–KPAN2–C17orf58*)[26]. The effect sizes of the latter three loci reproduced in the HGI cohort, but the associations were not significant, likely due to limited statistical power even in such a large cohort. This is a key challenge—it is notable that even with a population over eight times larger, the reported genome wide significant association between *FOXP4* and long COVID ($p = 1.76 \times 10^{-10}$)[25] did not reproduce in the 23andMe analysis.

Compared to the GWAS above with 6450 and 53,000 cases respectively, this study has only a very small case cohort available (n =413). It therefore lacks power to directly replicate many individual signatures, not least because it is very unlikely that the exact same high-order combinations of SNPs found in the original disease associated signatures will occur in sufficient numbers in such a small population to enable demonstration of statistical significance. We therefore focused on evaluating the reproducibility rates of the signatures that we had previously identified in the second dataset (i.e., the percentage that are also positively correlated with long COVID in All of Us) using a permutation testing approach.

Combinatorial analytics for complex diseases

Combinatorial analytics has been more successful than GWAS in identifying key genetic risk factors that capture the complex biology of similarly multifactorial and heterogenous diseases like ME/CFS, generating more mechanistic insights and reproducible findings across cohorts [27]. This lends itself to improved understanding of ancestry diversity, specific signature replication, and drug target validation, as well as new gene discovery.

The combinations of genetic variants ('disease signatures') identified by combinatorial analyses capture both the linear and non-linear effects of interactions between multiple genes. They can be used to identify individual patients who have specific disease signatures, enabling the identification of associations between the disease signatures associated with a mechanism and the symptoms presented by patients with those disease signatures. These can improve our understanding of complex diseases beyond the single SNP associations identified by GWAS [28, 29] and creates opportunities for clinically actionable diagnostic tests and the targeted trials of multiple drug repurposing candidates to provide clinical benefit to specific patient cohorts.

Aims of study

The PrecisionLife combinatorial analytics platform was previously used to identify disease signatures for Severe and Fatigue Dominant long COVID cohorts derived from the Sano Genetics' long COVID GOLD (Sano GOLD) study, and to highlight the biological similarities and differences between these two patient populations [30]. At the same time, combinatorial analysis was also undertaken on a General long COVID cohort encompassing all patients with a broader (and potentially less reliable) definition of the disease. The General cohort's results were not described in the original publication, which instead focused on the most well phenotyped and reliable cases.

The Severe cohort in this study was comprised of cases who self-reported the greatest variety and severity of symptoms, while the Fatigue Dominant cohort was comprised of cases who self-reported predominantly fatigue-associated long COVID symptoms. The study identified a total of 73 genes that were highly associated with at least one of these long COVID populations. Of these genes at the time of publication, 9 genes were linked to acute COVID-19, 14 genes were differentially expressed in a previous transcriptomic analysis of long COVID patients [31] and 9 genes were found that had been associated with ME/CFS in the previous combinatorial analysis of this disease[27].

In this study, we assessed the findings of all three of the original long COVID combinatorial analyses of the Sano GOLD cohorts in an independent, more ancestrally diverse patient population. We used genomic, clinical, and questionnaire data from the All of Us (AoU) population [32] to generate a long COVID cohort (using ICD-10 code U09.9) and evaluated the reproducibility of the findings from the original Sano GOLD study. We investigated the genes and mechanisms underlying the reproducible disease signatures, and evaluated the clinical phenotypes associated with each.

Materials and methods

Evaluating enrichment of reproducing long COVID disease signature in AoU cohort

We used a logistic regression approach to evaluate the disease association of previously identified disease signatures in the AoU study population. Individuals were coded as 1 if they possessed the exact combination of SNP genotypes comprising a signature and 0 if they did not. This term was included in the regression as an independent variable alongside covariates representing the top 5 genetic PCs (see Supplementary Table 1), with case–control status of the patients in the population (1 =case, 0= control) as the dependent variable.

The limited number of patients with ICD-10 codes for long COVID provides very limited power to replicate, i.e. statistically validate, individual disease signatures' disease associations in AoU, especially given the need for false discovery rate correction when testing the many signatures identified in the Sano GOLD dataset. Instead, we began by testing whether we observed a statistically significant enrichment of disease signatures that are also positively correlated with long COVID in the AoU cohort after controlling for population substructure.

For each of the three sets of disease signatures identified in the original Sano GOLD study (Severe, Fatigue Dominant, and General), we first counted the number of signatures where the logistic regression returns a positive coefficient (i.e., odds ratio >1) for the independent 'genetic signature' variable. Below we use the term 'reproducing' to denote signatures with odds ratio > 1 in the AoU test cohort, and 'reproducibility rate' to denote the percentage of tested signatures that have odds ratio > 1, as shown in Fig. 1.

Some of the original signatures could not be evaluated in AoU because one or more of their component SNP genotypes are not included in the dataset. These were excluded from the analysis (see Supplemental Table 2). Most of these missing SNPs are represented on the Illumina GDA array but these data were likely filtered out during the AoU dataset's QC processes. The distributions of signature frequencies in the Sano GOLD cohort for included and excluded signatures are shown in Supplemental Fig. 2. Excluded signatures derived from the Severe cohort were similar in frequency to the signatures tested in this study. Excluded signatures derived from the Fatigue Dominant and General cohorts tended to have lower frequencies in the Sano GOLD cohort than signatures that were tested in this study. This pattern likely reflects the fact that signatures with greater number of SNPs are more likely to include at least one SNP that is not present in the All of Us dataset and also tend to occur at lower frequencies in the population.

Many of the long COVID disease signatures are non-independent due to shared component SNP genotypes and linkage disequilibrium, preventing us from using standard statistical tests to evaluate the significance of our observed reproducibility rates. We therefore used a permutation-based approach to generate the expected distribution of observed reproducibility rates under the null hypothesis (i.e., no association between signatures and disease). We randomly shuffled the case–control vector 100 times, reran the logistic regression analysis for every signature and counted the number of disease signatures that have odds ratio greater than 1 for each random permutation. The *p*-value of the observed results is equal to the number of permutations in which the number of signatures with odds ratios above 1 is greater than or equal to the number of signatures with odds ratios above 1 in the original analysis.

From previous experience with other diseases, we have found that reproducibility rates for disease signatures are positively correlated with the frequency of the signature in the population. To test whether this is true for long COVID we filtered each of the three sets of disease signatures from the original analysis to generate subsets that occur in at least 4% or 5% of total cases in the AoU cohort (based on observations of reproducibility rates from prior unpublished studies in other diseases). We then reran the reproducibility analysis for each of these 'high frequency' subsets of signatures.

Finally, we evaluated the impact of signature complexity (i.e., number of SNP genotypes in the signature) on reproducibility rates. We split the set of signatures into sets comprised of 2, 3, 4, and 5 SNP genotypes and reran the analysis of reproducibility on each separately, with and without applying filtering by case-frequency.

Generation of long COVID cohorts

For this study, we identified a cohort of long COVID patients and matching controls from the AoU dataset

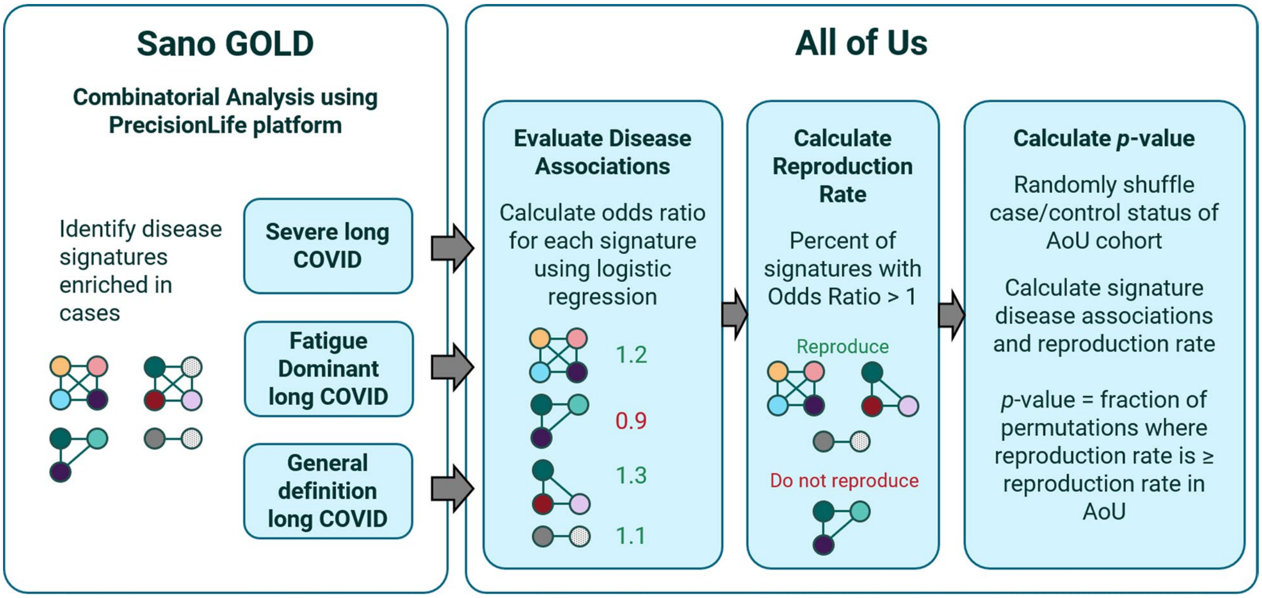
Sano GOLD
Combinatorial Analysis using PrecisionLife platform
Identify disease signatures enriched in cases
Severe long COVID
Fatigue Dominant long COVID
General definition long COVID
All of Us
Evaluate Disease Associations
Calculate odds ratio for each signature using logistic regression
1.2
0.9
1.3
1.1
Calculate Reproduction Rate
Percent of signatures with Odds Ratio > 1
Reproduce
Do not reproduce
Calculate p-value
Randomly shuffle case/control status of AoU cohort
Calculate signature disease associations and reproduction rate
p-value = fraction of permutations where reproduction rate is ≥ reproduction rate in AoU

Fig. 1 Study flow and overview reproducing disease signatures identified in the original combinatorial analysis of the Sano GOLD dataset[30] in a disjoint and more ancestrally diverse AoU cohort

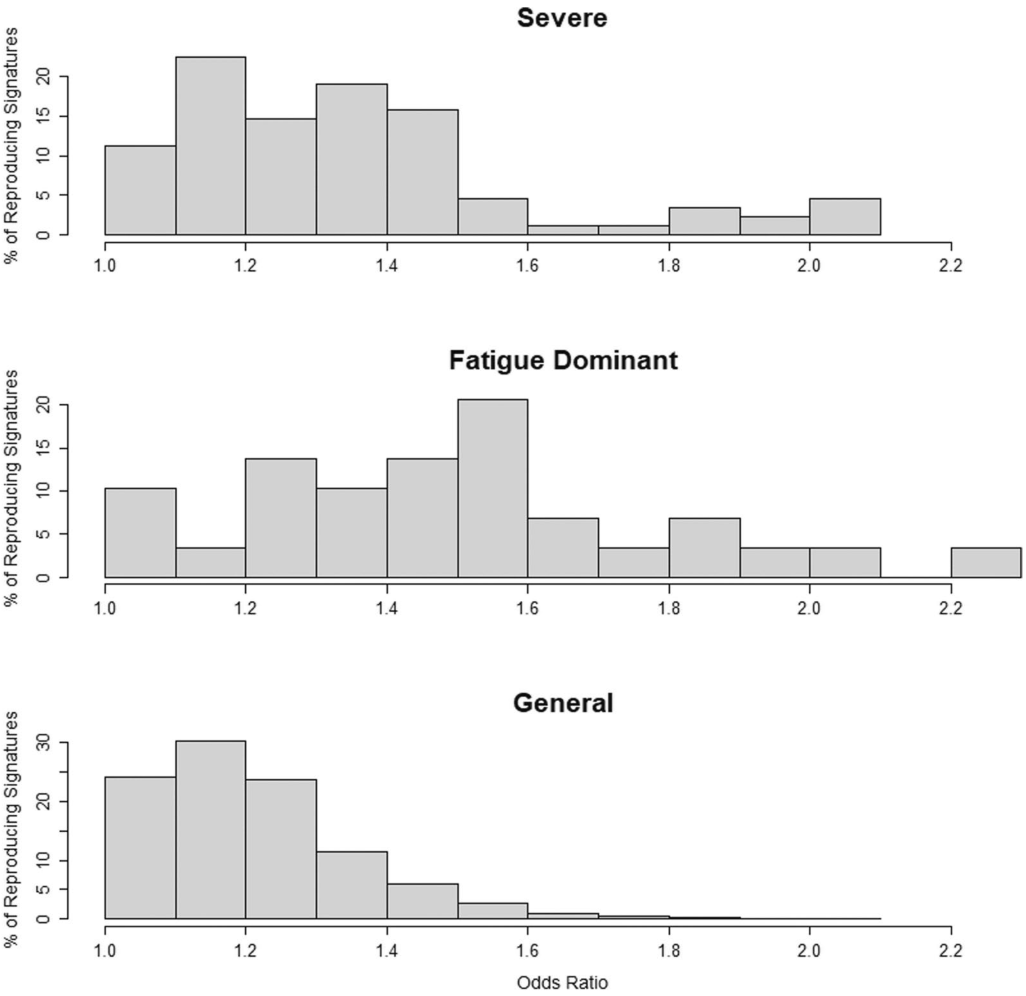

Fig. 2 Distribution of observed odds ratios in AoU for reproducing signatures with high case frequencies (> 5%) in the Severe (top), Fatigue Dominant (middle) and General (bottom) cohorts

(accessed on December 10th 2024). AoU provides data [33] for nearly 850,000 American participants, including genomic data derived from the Illumina Global Diversity Array (GDA) [34] (n =312,925), electronic health records (EHR, n= 254,700), health questionnaires (n =412,220), and COPE COVID-19 survey (n =100,220) [35]. The AoU dataset was designed to capture data for a diverse group of individuals, including non-European ancestry groups often underrepresented in genomic datasets, and the cohort selected for this study reflects this diversity.

The baseline long COVID cohort was created by selecting all 458 individuals with GDA genotyping data who have a diagnosis of long COVID, using ICD-10 code U09.9 (post-acute COVID-19). We note that this criterion, which implies a prevalence of long COVID less than 0.2%, almost certainly excludes many patients with long COVID based on published estimates of long COVID prevalence of between 6.9% and 14% [36–38].

The control cohort was generated by selecting individuals with GDA genotyping data who have evidence of SARS-CoV-2 infection, either based on a reported positive COVID-19 test in the COPE COVID-19 survey (n =3615) or presence of ICD-10 codes B97.21 or U07.1 (n =17,024). We excluded individuals with long COVID based on ICD-10 code U09.9 as well as any individual with a history of symptomatic phenotypes consistent with long COVID or other post-viral fatigue syndromes (see Supplemental Table 3). Applying these criteria, our maximum control population included 9774 individuals.

We used the sex-imputation functionality of PLINK [39] to identify the genetic sex of each of the individuals in the full GDA dataset. 2.9% of total samples could not be reliably identified as male or female and were excluded from the study. 57.6% were identified as female and 39.5% were identified as male, which broadly agrees with the self-reported distribution of sex at birth from the AoU demographics questionnaire (59% female, 39% male, 2% skip/unknown).

Case and control matching using stratified sampling

To create a balanced dataset and reduce potential confounding effects of population substructure, we created a subset of controls that match the demographic distribution (i.e., sex and self-reported race/ethnicity) of the long COVID cases. We used genetic sex inferred by PLINK (using the command –check-sex) as well the answers to the demographics survey on self-reported race and ethnicity to split the cohort into subgroups, and we compared the percentage of the baseline cases and potential controls that fall in each category. The results showed that some subgroups were over- or under-represented in cases vs controls, e.g. white, female, non-Hispanics accounted for 38.5% of cases but only 29.7% of controls.

We adjusted our long COVID cohort by removing all individuals whose sex was undetermined during PLINK sex-imputation. We also removed all individuals whose self-reported race and/or ethnicity was coded as "None of these/I prefer not to answer/PMI: Skip" as these demographics do not allow accurate 'matching' with the control population. This created a final long COVID case population of 413 individuals (see Table 1 for demographic distribution).

The set of potential controls allowed us to create a final study cohort with a 1:10 case:control ratio and similar demographic splits in the case and control sub-cohorts. We used a probabilistic function to apply a stratified sampling technique using granular subgroups based on three demographic values (as illustrated in Table 2) to the baseline potential controls and match the distribution of the demographic subgroups in the long COVID cases as closely as possible.

Prior to sampling, we also removed any age-based outliers from the control cohort (i.e., any individuals whose age was outside the range of ages represented in the long COVID case cohort). Additional information on the demographic breakdown of cases and controls, including prevalence of comorbidities is included in Supplemental Table 4. The concordance between the self-reported demographic data and the AoU genetic ancestry predictions[33] was very high (88.0–99.4% for matched groups) in the study cohort (Supplemental Table 5).

We used principal component analysis (PCA) to model any remaining population substructure within the AoU study cohort. We first removed all SNPs that are associated with the sex chromosomes or the MHC region on chromosome 6 or that have minor allele frequency less than 0.05. We then conducted LD-pruning in PLINK 1.9 (–indep-pairwise 50 5 0.2) before generating genetic PCs using the PLINK –pca command, as recommended elsewhere [40–42]. We selected the top 5 PCs for use in our analyses based on the associated eigenvalues (Supplemental Table 1).

Table 1 The distribution by genetic sex, self-reported race, and self-reported ethnicity of cases and controls in the final All of Us long COVID cohort

Demographic	Subgroup	Final cases % (n =413)	Baseline Controls % (n =8683)	Stratified Sampled Controls % (n =4130)
Genetic sex	F	69.7	65.2	68.3
Genetic sex	M	30.3	34.8	31.7
Self-reported race	Asian or none indicated	15.7	24.9	17.0
Self-reported race	Black or African-American	18.6	19.5	19.5
Self-reported race	White	65.6	55.6	63.5
Self-reported ethnicity	Hispanic or Latino	17.2	26.4	18.5
Self-reported ethnicity	Not Hispanic or Latino	82.8	73.6	81.5

Table 2 Two examples of how stratified sampling balances the frequency of granular subgroups (full breakdown is not available due to reporting restrictions imposed by AoU for rare subgroups)

Sex	Self-reported race	Self-reported ethnicity	Final cases (%)	Baseline controls (%)	Stratified sampled controls (%)
F	Black or African-American	Not Hispanic or Latino	14.5	13.0	14.7
F	White	Not Hispanic or Latino	42.6	33.2	40.2

Long COVID disease signatures

We previously identified long COVID associated disease signatures in two patient cohorts derived from the Sano GOLD study cohort[30] and a third unreported patient cohort using a broader definition of the disease (Supplemental Table 6). This resulted in:

1. 1188 signatures mapped to 43 genes, from a 'Severe' cohort of patients who reported the greatest variety and severity of long COVID symptoms.
2. 1435 signatures mapped to 35 genes, from a 'Fatigue Dominant' cohort of patients who reported predominantly fatigue-associated long COVID symptoms.
3. 6445 signatures mapped to 165 genes, from a 'General' cohort of patients who reported they were still suffering continuation or development of new symptoms 12 weeks after the initial SARS-CoV-2 infection, with these symptoms lasting for at least 2 months with no other explanation.

In contrast to the diverse AoU American study cohort, the Sano GOLD study cohort was comprised of British patients of predominantly white European ancestry (> 91% of the cohort), with Asian ancestry (~ 4%) as the largest non-white European demographic.

Ancestry-specific analyses

To test whether the observed rates of disease signature reproducibility apply broadly across traditionally underserved patient cohorts, we created three ancestry-specific sub-cohorts consistent with the demographic categories used to match cases with controls:

1. White—patients who self-identify as 'white'
2. Black/African-American—patients who self-identify as 'black' and/or 'African-American'
3. Hispanic/Latino—patients who self-identify as 'Hispanic', 'Latino', and/or 'Latina'

Note that these cohorts are not all mutually exclusive, as AoU includes separate questionnaire questions for self-reported race and Hispanic/Latino identification.

We again used genetic principal components to control for any indirect relationships between signature frequency and disease prevalence resulting from population substructure (including relatedness or broader shared ancestry between patients). We conducted separate PCAs for each sub-cohort dataset using the approach described above for the whole cohort. We then selected the first five PCs as covariates for each ancestry-specific analysis after confirming that they explained sufficient variance in the dataset (see Supplementary Table 1).

We restricted the ancestry-specific analyses to the sets of signatures that occur in more than 5% of cases. Given the small sample size and low statistical power for ancestry-specific sub-cohorts, it is most appropriate to assess differences in reproducibility rate for the sets of signatures that exhibit the strongest reproducibility statistics in the full cohort.

Evaluating enrichment of reproducing long COVID disease signature in AoU cohort

Finally, we tested whether any of the original disease signatures replicate, i.e., are individually significantly associated with long COVID in AoU. To minimize the FDR correction required for multiple tests, we restricted the analysis to the sets of high-frequency signatures that occur in more than 5% of cases. Output for the three original combinatorial analyses were assessed separately. Uncorrected *p*-values were obtained from the logistic regression with genetic PC covariates. 'Reproducing' signatures have *p*-values <0.05 after FDR correction via the Benjamini–Hochberg procedure [43]. We also assessed significance using the more conservative Bonferroni adjustment [44].

The SNPs in the disease signatures associated with long COVID in AoU were mapped to genes using an annotation cascade process against the human reference genome (GRCh38), as detailed in Das et al. (2022)[27]. SNPs located within the coding region of a gene (or genes) were mapped directly to the gene(s) and any remaining SNPs within 2kb upstream or 0.5kb downstream were mapped to the closest gene(s).

Sensitivity analysis for misphenotyped controls

The fraction of All of Us participants with an ICD-10 code for long COVID (0.2%) is much lower than published estimates of long COVID prevalence of between 6.9% to 14% [36–38], implying that some individuals coded as 'controls' in our analysis likely have or have had long COVID. We performed a sensitivity analysis to estimate the effects of these misphenotyped 'controls' on reproducibility rate.

We began by applying different misphenotyping rates (i.e., 5% or 10% of controls that should be coded as cases) and calculated the number of 'true' cases and controls in our cohort under each scenario. We then assumed a 'true' odds ratio and calculated the number of corresponding cases and controls for a signature with that odds ratio and a selected frequency in cases.

Next, we assumed that each of the 'true' cases has the same probability of being misphenotyped as a control based on the number of observed cases in our cohort vs.

the number of 'true' cases in our sensitivity analysis. This allowed us to use a binomial distribution to randomly assign a number of 'true' cases with the signature that were misphenotyped as controls. We then calculated an 'observed' odds ratio that adjusts for the number of misphenotyped patients with and without the signature. We repeated this random draw 1 million times to calculate the probability of observing non-replication (i.e., observed odds ratio ≤1) for a signature with the selected true odds ratio and case frequency under the two misphenotyping rate assumptions.

Results

Reproducibility of overall long COVID disease associations in AoU cohort

We were able to test 5343 of the 9068 long COVID disease signatures originally identified in the three Sano GOLD sub-cohorts. The untested signatures all contained at least one SNP genotype that was not present in the post-QC AoU genotype dataset. Of the tested signatures, 1766 occur in greater than 5% of cases and 3100 occur in greater than 4% of cases in AoU.

When we restricted the analysis to signatures with case frequency greater than 5%, we consistently observed a significant enrichment of signatures (77–83%, $p < 0.01$) that are positively associated with long COVID in the AoU cohort across all three sets of disease signatures (Table 3). As it is not possible to calculate very small probabilities with a permutation-based approach, *p*-values reported as <0.01 means that the observed reproduction was reported in 0/100 permutations.

Notably, the percentage of signatures with odds ratios greater than 1 in AoU is much larger than observed in any of the permutations where cases and controls were randomly assigned to patients (e.g., 82% vs. a maximum of 57% in the random permutations for the Severe cohort). This result confirms that many disease signatures are non-randomly associated with increased long COVID prevalence in AoU.

Overall reproducibility rates are lower when we apply a less stringent 4% frequency cutoff, but the enrichment is still highly significant (60–71%, $p < 0.01$). That is, the observed number of signatures with odds ratios greater than 1 in AoU exceeds the maximum number of signatures with odds ratio greater than 1 in the random permutations. We did not observe any significant enrichment of reproducing signatures when we included low-frequency signatures in our analysis.

The distributions of odds ratios for the reproducing high-frequency (> 5%) signatures are shown in Fig. 2. 89% and 90% of the reproducing signatures from the Severe and Fatigue Dominant studies respectively have odds ratios greater than 1.1 in AoU, while 17% and 48% have odds ratios greater than 1.5. The mean odds ratio for the Severe signatures is 1.35 and the maximum is 2.09, while the mean odds ratio for the Fatigue Dominant signatures is 1.49 and the maximum is 2.22. Thus, the reproducing disease signal from these studies largely represents signatures that are individually strongly associated with increased disease prevalence.

The reproducing signatures from the General study tend to have lower odds ratios than the other studies (Fig. 2). 75% of reproducing signatures have odds ratios greater than 1.1 in AoU and 5% have odds ratio greater than 1.5. The mean odds ratio is 1.21 and the maximum is 2.10. Thus, although these signatures included many with relatively weak disease associations, they also include signatures with strong effect sizes.

The relative enrichment of low odds ratios for signatures from the General study likely reflects the greater number of cases (Supplemental Table 6) and greater statistical power associated with the Sano GOLD study cohort. That is, the General study was better suited to detect signatures with lower effect sizes relative to

Table 3 Reproducibility statistics in AoU for long COVID disease signatures derived from three Sano GOLD sub-cohorts

Case Frequency Filter	Sub-cohort	# signatures tested	# (%) signatures with odds ratio > 1 in AoU	% signatures with odds ratio > 1 in permutations mean [range]	*p*-value
5%	Severe	109	89 (82%)	42% [31–57%]	**< 0.01**
	Fatigue dominant	35	29 (83%)	43% [23–71%]	**< 0.01**
	General	1622	1,249 (77%)	44% [30–59%]	**< 0.01**
4%	Severe	243	163 (67%)	41% [30–56%]	**< 0.01**
	Fatigue dominant	85	60 (71%)	43% [21–59%]	**< 0.01**
	General	2772	1,663 (60%)	43% [30–59%]	**< 0.01**
none	Severe	668	234 (35%)	43% [32–53%]	0.96
	Fatigue dominant	740	252 (34%)	46% [33–57%]	0.98
	General	3935	1,731 (44%)	42% [33–51%]	0.36

p-values reported as <0.01 means that the observed reproduction was reported in 0/100 permutations the smaller Severe and Fatigue Dominant cohorts. The weaker disease associations may also reflect the relative reliability of the criteria used to define the Sano GOLD cohorts, as we believe that the case definition criteria for the General cohort is less accurate than the criteria used to identify patients with Severe and Fatigue Dominant long COVID subtypes. An overlaid comparison of the odds ratio distributions for the three cohorts is also shown in Supplemental Fig. 1.

Reproducibility statistics are strongest for high case frequency (> 5%) signatures comprised of 4 or 5 SNP genotypes, as measured both by percent reproducing (i.e., odds ratio > 1) and *p*-value (Table 4). Notably, across all three analyses, roughly twice as many 4- and 5- SNP signatures have odds ratios greater than 1 in AOU than would be expected due to random chance based on the mean reproducibility rates for the random permutations.

We similarly observed that reproducibility statistics are strongest for higher complexity signatures when applying a frequency cut-off of 4% (Supplemental Table 7). We observed no clear association between signature complexity and reproducibility rates for low frequency signatures (Supplemental Table 8).

To ensure that the sets of long COVID disease signatures are broadly reproducible across patients, we conducted separate analyses for self-reported white, black/African-American, and Hispanic/Latino sub-cohorts (Table 5).

We observed a highly significant enrichment of positively correlated disease signatures among

Table 4 Reproducibility statistics by signature complexity (i.e., number of SNP genotypes comprising disease signatures) in AoU for high case frequency (> 5%) long COVID disease signatures derived from three Sano GOLD sub-cohorts

Signature Complexity	Sub-Cohort	# signatures tested	# (%) signatures with odds ratio > 1 in AoU %	% signatures with odds ratio > 1 in permutations mean [range]	*p*-value
2	Severe	1	1 (100)	40% [0–100%]	0.43
	Fatigue Dominant	6	4 (67)	47% [0–100%]	0.29
	General	14	6 (43)	45% [7–86%]	0.58
3	Severe	20	11 (55)	47% [5–80%]	0.34
	Fatigue Dominant	16	13 (81)	43% [13–88%]	**0.01**
	General	784	596 (76)	44% [22–67%]	**< 0.01**
4	Severe	30	27 (90)	41% [23–60%]	**< 0.01**
	Fatigue Dominant	7	6 (86)	42% [0–100%]	**0.05**
	General	325	267 (82)	44% [26–58%]	**< 0.01**
5	Severe	58	50 (86)	42% [22–60%]	**< 0.01**
	Fatigue Dominant	6	6 (100)	42% [0–100%]	**0.01**
	General	501	386 (77)	43% [30–53%]	**< 0.01**

p-values reported as < 0.01 means that the observed reproduction was reported in 0/100 permutations

Table 5 Reproducibility statistics in AoU for high case frequency long COVID disease signatures (> 5% of cases) derived from three Sano GOLD sub-cohorts, broken down by self-reported ancestry

Self-reported Ancestry	Sub-Cohort	# signatures tested	# (%) signatures with odds ratio > 1 in AoU %	% signatures with odds ratio > 1 in permutations mean [range]	*p*-value
White (271 cases)	Severe	109	85 (78)	46% [33–61%]	**< 0.01**
	Fatigue Dominant	35	30 (86)	46% [20–71%]	**< 0.01**
	General	1622	1217 (75)	48% [34–65%]	**< 0.01**
Black/African American (77 cases)	Severe	109	62 (57)	47% [31–62%]	0.06
	Fatigue Dominant	35	18 (51)	45% [20–66%]	0.33
	General	1622	908 (56)	46% [33–59%]	**0.05**
Hispanic/Latino (71 cases)	Severe	109	72 (66)	48% [32–61%]	**< 0.01**
	Fatigue Dominant	35	20 (57)	46% [23–80%]	0.15
	General	1622	876 (54)	47% [35–60%]	0.14

p-values reported as < 0.01 means that the observed reproduction was reported in 0/100 permutations self-reported white patients. This result confirms that the observed enrichment of reproducible disease associations in the all-ancestry cohort does not simply reflect population substructure in the dataset (i.e., indirect correlations between disease prevalence and signature frequency that arise due to shared correlations with ancestry).

Reproducibility rates were lower in the self-reported black/African-American and Hispanic/Latino sub-cohorts relative to the self-reported white sub-cohort, but consistently above the mean values observed in the random permutations. Two of the observed enrichment values were statistically significant ($p < 0.05$) despite the very small number of cases (71 and 77) and consequent weak statistical power in these sub-cohorts.

More than 85% of the long COVID genes identified across the three Sano GOLD long COVID cohorts mapped to one or more disease signatures that have >4% case frequency and were also positively associated with long COVID in the AoU cohort (see Table 6, Fig. 3). Out of the 73 genes identified in the Sano GOLD long COVID study[30], 15 genes could not be tested due to missing SNPs in the AoU dataset. 76% (44/58) of the remaining genes map to disease signatures that reproduced in AoU. These genes are linked to a wide range of biological processes and mechanisms including dysregulated immune response and metabolic pathways, development of chronic inflammation, and cognitive dysfunction.

Of the 13 repurposing gene candidates identified in the Sano GOLD study, 11 (85%) map to at least one disease signature that reproduces in AoU (see Supplemental Table 9). These genes include *TLR4* which has been shown to protect against long-term cognitive impairment pathology caused by SARS-CoV-2[45]. Inhibition of TLR4 in a mouse model was shown to prevent long term cognitive pathology including synapse elimination and memory deficits that are caused by the SARS-CoV-2 Spike protein. Previous clinical studies have shown that antagonizing TLR4 signaling has the effect of dampening the pathological cytokine storm observed in patients with severe acute COVID-19 and reduces mortality rates in hospitalized COVID-19 patients [46, 47].

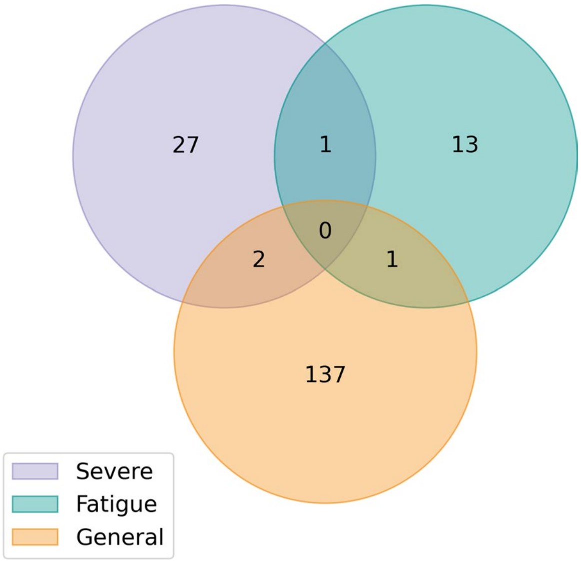

Fig. 3 Venn diagram showing overlap of genes identified in three Sano GOLD sub-cohorts (Severe, Fatigue Dominant and General) that were positively correlated with long COVID in AoU

The results of the misphenotyping sensitivity study are shown in Supplemental Table 10. For signatures with 5% frequency, we estimate that the presence of long COVID patients among the controls in our dataset will cause the observed reproduction rates to decrease by 3%–5% for signatures with true odds ratio =1.5, by 18–20% for signatures with true odds ratio =1.2, and by 27–39% for signatures with true odds ratios =1.1, depending on the frequency of misphenotyping. The probability of failing to observe reproduction due to misphenotyped cases further increases for lower frequency signatures with moderate-to-strong odds ratios.

Table 6 Reproducibility statistics in AoU for genes associated with high case frequency (> 4% and > 5%) long COVID disease signatures identified in three Sano GOLD sub-cohorts

Sub-Cohort	# Genes identified in Sano GOLD cohort	# Genes evaluated in AoU	# Genes mapped to signatures with > 4% case frequency	# (%) of Genes mapped to signatures with > 4% case frequency that have odds ratio > 1 in AoU %	# Genes mapped to signatures with > 5% case frequency	# (%) of Genes mapped to signatures with > 5% case frequency that have odds ratio > 1 in AoU
Severe	43	35	35	30 (85.7)	25	22 (88.0)
Fatigue Dominant	35	28	16	15 (93.8)	6	5 (83.3)
General	165	151	150	140 (93.3)	138	128 (92.8)

We note that the uncertainty due to misphenotyping can also result in observed false positives (i.e., signatures that have true odds ratios < 1 in a properly phenotyped cohort but that exhibit observed odds ratios > 1). However, the expected probability of false positive 'replication' of non-causal signatures due to misphenotyping does not exceed 50% for non-rare signatures and therefore cannot explain the observed results.

Replication of individual disease signatures in All of Us

The above analyses focused on demonstrating an overall enrichment of disease signatures and genes that are positively correlated with long COVID in AoU, recognizing that the small size of the AoU cohort severely limits wide-scale replication. To achieve sufficient power to statistically validate individual signatures, we limited our replication analysis to the subset of signatures with case frequencies above 5%.

Four high-frequency disease signatures from the Severe Sano GOLD analysis were significantly associated with increased prevalence of long COVID in AoU, one of which was still significant after applying the more conservative Bonferroni FDR correction (Table 6). All four signatures are comprised of five SNP genotypes, each of which contributes to the overall association with disease in AoU (i.e., removing any of the SNP genotypes from the signature results in a lower odds ratio). This observation highlights the utility of the combinatorial analysis approach for identifying genetic disease associations.

Two of the replicating disease signatures from the Severe analysis mapped to the gene *CCDC146* and one mapped to *D2HGDH*. These genes have different functions and affect different potential mechanism of action hypotheses for their role in the development of long COVID. CCDC146 is a ubiquitous centriole and microtubule-associated protein linked to cognitive functioning and type 2 diabetes [48]. D2HGDH is an enzyme involved in mitochondrial functioning, that also exhibits anti-inflammatory effects [49].

Two disease signatures from the Fatigue Dominant Sano GOLD analysis were significantly associated with increased prevalence of long COVID in AoU, one of which was still significant after applying the more conservative Bonferroni FDR correction (Table 6). The latter is comprised of two SNP genotypes, while the other is comprised of five SNP genotypes. Each of the individual SNP genotypes contribute to the signatures' association with disease in AoU.

None of the signatures from the General cohort in the Sano GOLD analysis were significantly associated with increased prevalence of long COVID in AoU. Although this output includes the signature most strongly associated with long COVID (by uncorrected *p*-value), it does not survive the stringent FDR correction for the large number of signatures from this analysis.

Finally, if we pool the signatures from the Severe and Fatigue Dominant cohorts into a single analysis (excluding the large number of signatures from the General cohort to avoid the need for stringent FDR correction), then 5 of signatures in Table 7 remain significant under the combined Benjamini–Hochberg FDR correction. These include all four significant signatures from the Severe analysis and the top signature from the Fatigue Dominant analysis.

Discussion

Studies performed using traditional GWAS and meta-GWAS approaches on large patient populations (6,450 cases and 53,764 cases) respectively identified a single locus and three loci associated with long COVID[25,26], although there was no statistical replication of the findings between these studies.

The original combinatorial analysis of Sano's GOLD cohort identified 9,068 genetic disease signatures and 73 genes that were significantly enriched in two small UK-based long COVID patient cohorts (Severe n_cases =459 and Fatigue Dominant n_cases =477)[30]. In this original analysis, 28/43 genes found in the Severe cohort were also significantly associated with disease in the Fatigue Dominant cohort, and 25/35 genes from the original Fatigue Dominant analysis were also associated in the Severe cohort. 25 genes (15 from Severe and 10 from Fatigue Dominant) were found to be unique to those cohorts.

92% of the genes and 60–83% of the medium/high-frequency disease signatures from the Sano GOLD results that are also represented in the AoU dataset were positively correlated with long COVID in this independent US-based population. For disease signatures that occur in at least 5% of patients, between 77 and 83% were positively correlated with long COVID prevalence in both the Sano GOLD and AoU cohorts, far more than we would expect to randomly observe if the signatures were uncorrelated with disease biology. Although we defined a 'reproducing' signature as one that has any odds ratio greater than 1, most reproducing signatures have relatively large odds ratios in AoU, indicating a strong association with increased disease prevalence.

At least five of the disease signatures found in Sano GOLD were individually significantly associated with increased prevalence of long COVID in the AoU population. The significant enrichment of positively-associated disease signatures further confirms that many additional signatures are non-randomly

Table 7 Replicating disease signatures that exhibit statistically significant associations with long COVID in AoU using Bonferroni-Hochberg FDR procedure

Signature	Gene	# Cases	# Controls	Odds ratio	*p*-value
Severe cohort					
(rs17035343 T/T) (rs1872513 C/C) (rs9312595 A/G) (rs4936114 A/G) (rs12454570 C/C)	*NBPF21P* (–) *ASB5* *PPP1R10P1* (–)	30	149	2.09	**0.0004**
(rs17035343 T/T) (rs1872513 C/C) (rs9312595 A/G) (rs58438895 A/A) (rs4936114 A/G)	*NBPF21P* (–) *ASB5* *CCDC146* *PPP1R10P1*	29	144	2.09	0.0005
(rs17035343 T/T) (rs1872513 C/C) (rs9312595 A/G) (rs1109968 A/A) (rs4936114 A/G)	*NBPF21P* – *ASB5* *CCDC146* *PPP1R10P1*	29	144	2.09	0.0005
(rs6716743 G/G) (rs2010874 T/T) (rs13096228 A/G) (rs67844017 A/A) (rs79853277 T/T)	*D2HGDH* *LSAMP* (–) *LINC02520* (–)	28	136	2.15	0.001
Fatigue dominant cohort					
(rs9515203 C/C) (rs11633336 A/A)	*COL4A2* *SLC12A1*	27	120	2.33	**0.0004**
(rs10914896 G/G) (rs10229643 A/A) (rs9960341 A/A) (rs2076584 C/C) (rs17702926 C/C)	(–) *GLCCI1* *THEMIS3P* *RIN2* *AL024495.1*	23	117	2.02	0.003

p-values in bold are also significant under Bonferroni FDR correction. Signatures from each cohort analysis were evaluated separately. Odds ratios reflect the number of total case and controls with genotype data for all component SNPs, which differs between signatures. Genes mapped to the SNPs in the signatures are listed in the same order as the SNPs and they are shown as (–) if a SNP could not be mapped to any gene

associated with disease but cannot be individually validated due to the very low statistical power provided by the small number of long COVID patients in the dataset (n =413). Together these results demonstrate a significant enrichment and reproduction of disease signal, broadly validating the results of the original analysis.

Replication of individual loci was not shown by the previous, much larger GWAS studies. The fact that five disease signatures did directly replicate, even with such a small sample size, suggests that if we were able to analyze cohorts closer to the size of the GWAS studies, it is likely that many more signatures would be directly replicated. As the All of Us long COVID cohort increases in size we will revisit this analysis, which will allow us to estimate the sample size needed for this broader replication.

Importantly, the results provide strong supporting evidence for a much broader range of genetic associations with long COVID than has been uncovered by GWAS studies to date. This provides evidence highly consistent with a strong biological basis of the disease and the hypothesis that patients' genetics influence their susceptibility to developing long COVID (and their predominant symptoms) following recovery from acute SARS-CoV-2 infection.

The AoU ancestry distribution differs significantly from the mainly (> 91%) white British patient cohort used in the original combinatorial analysis. Disease signature reproducibility rates are very strong in the sub-cohort of self-identified white patients, as expected given the similarity in ancestry between that cohort with the original Sano GOLD dataset. Signature reproducibility rates are lower in sub-cohorts of self-identified black/African-Americans and Hispanic/Latinos, but we still observe significant enrichment of disease signatures despite very small sample sizes.

This therefore represents the first reproduction of long COVID genetic associations across multiple populations with substantially different ancestry distributions. Given the degree of reproducibility of results across diverse populations, these findings may have broad clinical application which could promote better health equity.

The lower signature reproducibility rates among the self-identified black/African American and Hispanic/Latino sub-cohorts relative to the self-identified white sub-cohort could be caused by a combination of the lower sample sizes for these cohorts, differences in allele frequencies, and differences in epidemiological and/or environmental influences between the ancestry groups.

For example, the disease signatures identified in the Sano GOLD cohort may have reduced effect sizes in non-white European cohorts due to an increased frequency of 'actively protective' signatures in those populations, i.e., one or more SNP genotypes that wholly or partially mitigate the disease associations of a set of 'causative' disease signatures [50]. The combinatorial analysis in this study focused only on causative disease signatures and did not include any analysis of protective signatures[30]. Incorporating actively protective features into the set of disease signatures should increase their predictivity for identifying 'high-risk' patients and improve reproducibility statistics.

These results highlight a pressing need to identify large, diverse, well-phenotyped cohorts of long COVID patients. Many long COVID specific datasets such as Sano GOLD are comprised predominantly of patients with white European ancestry. In contrast, All of Us includes a highly diverse patient cohort, but lack of reliable data identifying which participants have a history of long COVID prevents us from reliably obtaining sufficient sample sizes to conduct a combinatorial analysis aimed at identifying novel disease signatures.

Having access to larger and more diverse populations with a confirmed diagnosis is essential to enabling primary analysis of disease risk and protective factors within these ancestry cohorts and adding to our understanding of the factors underpinning disease in those populations. In turn this would also help us build more inclusive and transferrable disease risk models. Notably, combinatorial analysis of diverse long COVID patient cohorts could potentially identify disease signatures that were not detected in predominantly white European cohorts due to low relative case frequencies, but which have greater frequency and importance for disease biology in other patient cohorts. Such signatures could be used to better estimate patients' relative susceptibility to developing long COVID.

Evaluating the output of the PrecisionLife combinatorial analysis platform

We observed high rates of reproducibility among disease signatures derived from all the Sano GOLD cohorts and showed that these rates of disease signature reproducibility were strongly correlated with the frequency of signatures in the original study cohort. We observed slightly higher overall rates of reproducibility in the Severe and Fatigue Dominant cohorts which have fewer high case frequency disease signatures relative to the broader 'General' long COVID cohort.

Rates of reproducibility were highest for disease signatures comprised of four or five SNP genotypes, suggesting that combinatorial genetic interactions play an important role in the biology of long COVID. This also provides supporting evidence for the combinatorial analytic approach's ability to detect a broad and clinically informative set of genetic disease associations in otherwise intractable complex diseases.

The predictive value of common vs rare signatures

In contrast to these mid/high case frequency signatures, when analyzing the entire set of disease signatures from the original analyses including low frequency signatures, only 34–44% were consistently correlated with long COVID prevalence. This implies that rarer signatures may replicate between populations more poorly, an observation that is consistent with similar findings in GWAS and polygenic risk score studies [51–55]. There are several explanations for this observed correlation between signature frequency and reproducibility rates.

First, statistical power is proportional to sample size, which is already limited in the reproducibility analysis due to the very small number of confirmed long COVID patients in AoU. Signatures with frequencies below 5% are expected to occur in 21 or fewer AoU cases. This small sample size results in large variance in expected rates of reproducibility under the null model and a high probability of observing odds ratios less than one due to random sampling even when signatures are biologically relevant to disease.

Second, we demonstrated that reproduction rates for low frequency signatures with moderate-to-high true odds ratios are most strongly negatively affected by the likely inclusion of patients with long COVID in the control cohort.

Third, due to the high case:control skew (1:10) in our dataset, rare signatures were often more likely to be negatively correlated with disease under the null model. In the most extreme scenario, a signature that occurs in one person in the dataset is 10 times more likely to randomly occur in a control (resulting in a negative odds ratio) than a case (resulting a positive odds ratio). This bias caused the mean numbers of signatures that randomly exhibit odds ratios above 1 in the null model permutations to range between 41 and 46% - below the 50% expectation for a balanced dataset.

Fourth, rare signatures appeared to be more reflective of subpopulation structure in the original Sano GOLD dataset. Including genetic principal components as covariates resulted in 4% fewer high-frequency signatures (i.e., those that occur in >5% of total cases) that are positively correlated with long COVID, relative to a logistic regression that did not include covariates for population substructure. In contrast, including genetic principal components in the analysis resulted in 52%

fewer replicating low-frequency signature (i.e., those that occur in < 4% of total cases).

Finally, more complex disease signatures (i.e., those comprised of 4 or 5 SNP genotypes) generally occur at lower frequencies in the population simply because there are more possible genotype combinations for a larger set of SNPs. The risk of overfitting to a dataset is known to increase with tree depth when applying tree-based machine learning algorithms [56] and the same potentially holds true for higher layer disease signatures derived from a layer-based mining approach. Applying a frequency filter therefore potentially mitigates the impacts of false positive SNPs by removing higher-order signatures.

We found no evidence, however, that increased signature complexity was associated with reduced reproducibility among high-frequency signatures. Rather, overall reproducibility rates were highest for 4-SNP and 5-SNP signatures relative to the small number of 2-SNP signatures. We also did not observe a correlation between signature complexity and reproducibility rate among low-frequency signatures. These results suggest that outputs of the combinatorial analyses of the Sano GOLD cohorts were not excessively overfitted to the original datasets and that presence of any false positive component SNP genotypes does not significantly affect the overall association with disease.

Although the results of this analysis suggest that false positive component SNP genotypes do not have a major effect on signature reproducibility, we could potentially improve the effect sizes and predictivity of these signatures by using AoU to further refine the set of signatures. This step entails testing each signature individually and removing any component SNP genotype that does not enhance the signature's association with disease in AoU. We have not included any refinement analysis in this study because it can potentially overfit the new set of signatures to the training dataset (AoU). A third cohort of long COVID patients would be required to properly evaluate the improvement in disease signature reliability that results from this refinement process.

Limitations of analysis

Reliably identifying which patients in AoU have a history of long COVID is currently challenging. We relied on ICD-10 coding, which is known to be inconsistently and inaccurately applied, to identify known cases. As noted above, published estimates of long COVID prevalence in the United States range between 6.9% to 14%, yet fewer than 0.2% of individuals in AoU have ICD-10 codes associated with long COVID.

This suggests that many long COVID patients have not been assigned the appropriate ICD-10 code. As a result, more than 10% of the controls in our AoU study cohort potentially could represent misclassified cases with unreported long COVID. We have, however, noted in similar studies of highly heterogenous diseases that constraining diagnostic criteria in this manner (at the cost of potentially missing true cases) is more effective than including a larger number of poorly diagnosed (potentially true negative) patients into the case group.

This type of phenotypic misclassification in datasets will generally weaken the observed effect sizes by artificially inflating the similarity between cases and controls [57]. This behavior is potentially problematic for reproducibility analyses, as the dilution of signal decreases the statistical power of the analysis [58]. For example, phenotypic misclassification increases the probability that a signature that is biologically correlated with increased disease risk will nonetheless exhibit an odds ratio less than 1 due to random sampling effects.

We therefore expect that the high degree of phenotypic misclassification in our dataset will have worked to reduce the overall rates of observed signature reproducibility. As such, the reproducibility statistics presented in this paper probably represent a low-end estimate of the true reproducibility rate. Better diagnostic criteria and the use of harmonized surveys of patients' self-reported symptoms would help pool patient datasets and compare results across cohorts.

We noted that 41% of signatures (3725/9068) could not be tested due to missing SNPs in AoU. These were predominantly SNPs that were removed from the AoU dataset before its release, presumably due to data quality issues [59]. The removal of these SNPs skew the reproducibility analysis toward slightly more common signatures, as signatures comprised of many SNPs tend to have lower frequencies in the population and are also more likely to include a missing SNP. Because they are enriched in rarer signatures, inclusion of these missing SNPs may have resulted in slightly weaker overall reproduction rates. They would not be expected to have a significant effect on the reproduction statistics for more common variants, however, and may have resulted in stronger p-values due to larger sample sizes of signatures tested.

There was much more heterogeneity and diagnostic uncertainty about the General Cohort in the original Sano GOLD analysis relative to the more well-defined Severe and Fatigue Dominant cohorts. This included small sets of long COVID 'patients' self-reporting no prior COVID-19 infection, and others reporting improvements in their health post COVID-19 infection.

It was therefore unsurprising that this more heterogeneous cohort was associated with more signatures in the Sano GOLD study, consistent with the assumption that a wider variety of phenotypes will be linked to more genes. This large number of signatures is problematic when attempting to statistically validate the replication of individual disease associations from the General cohort due to the need for more stringent FDR correction. As a result, none of the individual General cohort signatures statistically replicated, even though these include signatures that have stronger p-values than the signatures in the Fatigue Dominant and Severe cohorts, which did statistically replicate. A much larger cohort is required to overcome this multiple testing burden and allow for statistical validation of individual signatures from the General cohort.

Applications for healthcare

The identification of a set of genetic signatures that are consistently associated with increased prevalence of long COVID offers many opportunities for improving treatment of this poorly understood but highly prevalent and debilitating disease.

Firstly, although we have insufficient power to validate the full set of individual signatures in AoU, demonstrating that reproducing signatures are significantly enriched in a second dataset provides important confirmatory evidence of the original findings of the combinatorial analytics approach. To provide insights into potential drug therapies for long COVID, we further tested whether the signatures associated with novel drug targets and their related drug repurposing candidates are significantly correlated with increased long COVID prevalence in AoU. 27/30 (90%) of the genes represented in the > 5% disease signatures and 11 out of the 13 drug repurposing candidates identified in the original study were reproduced in this study. This lends weight to their prioritization in clinical efficacy trials especially for those with generic drugs.

Controlled open-label studies of similar design to the RECOVERY trial in Covid-19, which rapidly identified dexamethasone as an effective frontline therapy [60], can be undertaken on this set of generic drugs, benefiting from the additional evidence that one or more selected therapies is more likely to help the subset of patients who have those mechanisms' disease signatures in their genetic makeup.

Secondly, we can use the insights into disease biology that are reflected by the reproducing disease signatures to construct a combinatorial risk score that evaluates an individual patient's relative genetic susceptibility towards developing long COVID. Although genetic risk scores are not strictly diagnostic, especially in a pathogen triggered disease, they have substantial potential to be used by physicians for differential triage, i.e., to rapidly gauge the relative likelihood of different diagnoses when presented with ambiguous or indistinct symptoms and refer patients and/or select treatment options accordingly. As the utilization of large-scale COVID-19 testing fades, alternative tests that can help differentiate patients with long COVID from patients with other illnesses with similar symptoms will become increasingly useful in healthcare settings.

Constructing a combinatorial risk score from disease signatures is a more complex challenge than a conventional polygenic risk model—the latter assumes that all features (SNPs) act independently, whereas combinatorial disease signatures are often inherently correlated due to the sharing of SNP genotypes. Machine learning approaches can disentangle this complexity and non-independence and combine features such as disease signatures and their component SNP genotypes into a single predictive model. Although the small sample size of the AoU dataset is sufficient to train a combinatorial risk score using machine learning, a third (currently unavailable) independent dataset would be required to properly evaluate the relative increase in long COVID prevalence between subsets of patients flagged as having high and low genetic susceptibility.

Finally, the set of replicating disease signatures can be used to mechanistically stratify patients based on the causative etiologies most likely associated with their form of long COVID. This first entails assigning disease signatures to one or more mechanism-of-action (MoA) cluster(s) based on the gene(s) associated with the component SNP genotypes. We can then assess whether a patient has a significant excess or lack of disease signatures associated with a given MoA relative to the distribution of signature counts in the larger long COVID community. In essence, this mechanistic stratification tool is comprised of multiple combinatorial risk scores, each for a different set of mechanism related disease signatures. This can provide insight in the clinic into the selection of therapies that are matched to a patient's personal genetic makeup.

Unlike standard risk scores, which can be used to inform public health applications but provides more limited utility for personalized precision medicine [61], a mechanistic stratification tool would potentially ultimately enable healthcare practitioners to identify individualized treatment regimens including single- or multi-drug therapies that are most likely to generate a positive outcome for a given patient. In the case of long COVID these mechanistic insights also have other potential applications, as the Taylor et al. (2023) combinatorial analysis also found evidence for substantial overlap in disease biology between long COVID and myalgic encephalitis/chronic fatigue syndrome (ME/CFS)[30].

Conclusion

The level of reproducibility of results from the original Sano GOLD long COVID study in the All of Us population to the extent demonstrated is highly encouraging for the study of long COVID and other similarly complex diseases. These findings redefine our understanding of long COVID by uncovering a broad spectrum of reproducible genetic signatures, laying the foundation for new diagnostic innovations and targeted therapies that have the potential to revolutionize care for millions suffering from this debilitating condition worldwide.

The study demonstrates the level of reproducibility of results achievable using combinatorial analysis, even across very small populations with diverse ancestries in highly heterogenous diseases. Increasing reproducibility across patients with different ancestries is critically important for improving equitable representation and access to healthcare solutions. All of these studies would nonetheless obviously benefit from larger datasets with a wider population diversity, more secure diagnosis, more harmonized health/symptom surveys and deeper genomic, longitudinal clinical, immunological and metabolic data.

The results provide further compelling evidence for the detailed description of the genetic components of long COVID's complex disease biology that was presented in the original combinatorial analysis study[30]. We hope that this will provide confidence to explore some of these mechanisms and drug targets and help advance research into novel ways to diagnose the disease and accelerate the discovery and selection of better therapeutic options, both in the form of newly discovered drugs and/or the immediate prioritization of coordinated investigations into the efficacy of repurposed drug candidates.

We also hope that these findings will better establish a stronger appreciation of the role of genetic contributions to the etiology and lived experience of disease in long COVID patients and prove its underlying biological basis to the clinical community.

For the first time, a definitive test for the disease would enable clinicians to rapidly and accurately identify and triage patients, ensuring they receive timely and equitable access to care, and reducing the potential for misdiagnosis. It would also establish a definitive framework for measuring the public health impact of the disease, informing health policy and helping strategically prioritize research initiatives to make more rapid progress in addressing this massive global challenge and improving patients' lives.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-06535-x.

Supplementary material 1

Acknowledgements

Research described in this article has been conducted using data from the All of Us Research Program Controlled Tier Dataset (CDR) version 7 and Sano Genetics' Long COVID GOLD study. We gratefully acknowledge All of Us and Sano Genetics' Long COVID GOLD participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program and Sano Genetics for making available the participant data examined in this study. Special thanks to Gert Møller and Claus Erik Jensen, who initially developed the combinatorial analytics methodology, and the rest of the PrecisionLife, Metrodora and CODA teams.

Author contributions

SG, AR, MS, JS, SD, RG and KT contributed to the design of the study. JS designed the reproducibility analyses. JS, MP, SD and KC performed the analyses described in this manuscript. MP and KC conducted the analyses on the Researcher Workbench. All authors contributed to writing the manuscript and consent to publication.

Funding

The project was funded entirely by PrecisionLife Ltd and the Complex Disorders Alliance.

Availability of data and materials

Only data from existing All of Us and Sano GOLD study cohorts were analyzed and no new source data were collected for this study. Aggregate-level data for the All of Us cohort is publicly available at https://databrowser.researchallofus.org/ (Public Tier dataset). Individual-level data for the All of Us cohort, available in the Controlled Tier Dataset version 7, can be analyzed by approved researchers on the Researcher Workbench.

Declarations

Ethics approval and consent to participate

The Sano GOLD study has approval from the Wales Research Ethics Committee (REC) (IRAS 291221). Consent to participate has been received from all participants. Institutional Reviewing Board (IRB) approval was obtained prior to enrollment of patients in the All of Us Research Program. Informed consent for all participants is conducted in person or through an eConsent platform that includes primary consent, HIPAA Authorization for Research use of EHRs and other external health data, and Consent for Return of Genomic Results. The protocol was reviewed by the Institutional Review Board (IRB) of the All of Us Research Program (IRB Approval Date: Dec 03, 2021). The All of Us IRB follows the regulations and guidance of the NIH Office for Human Research Protections for all studies, ensuring that the rights and welfare of research participants are overseen and protected uniformly. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers (OT2 OD026549; OT2 OD026554; OT2 OD026557; OT2 OD026556; OT2 OD026550; OT2 OD 026552; OT2 OD026553; OT2 OD026548; OT2 OD026551; OT2 OD026555); Inter agency agreement AOD 16037; Federally Qualified Health Centers HHSN 263201600085U; Data and Research Center: U2 C OD023196; Genome Centers (OT2 OD002748; OT2 OD002750; OT2 OD002751); Biobank: U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: U24 OD023163; Communications and Engagement: OT2 OD023205; OT2 OD023206; and Community Partners (OT2 OD025277; OT2 OD025315; OT2 OD025337; OT2 OD025276). Results reported are in compliance with the

All of Us Data and Statistics Dissemination Policy disallowing disclosure of group counts under 20 to protect participant privacy.

Competing interest

AR is an employee of the Complex Disorders Alliance, SG and RG are co-chairs of the Complex Disorders Alliance's Scientific Advisory Board. JS, SD, KT, KC, MP and SG are employees of PrecisionLife Ltd. S.G. is a shareholder of PrecisionLife, Ltd.

Author details

[1]PrecisionLife Ltd., Unit 8b Bankside, Hanborough Business Park, Long Hanborough OX29 8LJ, UK. [2]Metrodora Institute, 3535 South Market Street, West Valley City, UT 84119, USA. [3]Complex Disorders Alliance, 2299 Summer St. #1140, Stamford, CT 06905, USA.

Received: 11 February 2025 Accepted: 24 April 2025
Published online: 08 May 2025

References

1. Al-Aly Z, Davis H, McCorkell L, et al. Long COVID science, research and policy. Nat Med. 2024;30:2148–64. https://doi.org/10.1038/s41591-024-03173-6.
2. Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol. 2023;21(3):133–46.
3. Blitshteyn S, Verduzco-Gutierrez M. Long COVID: a major public health issue. Am J Phys Med Rehabil. 2024;10:10–1097.
4. Callard F, Perego E. How and why patients made Long Covid. Soc Sci Med. 2021;268: 113426.
5. Davis HE, Assaf GS, McCorkell L, Wei H, Low RJ, Re'em Y, Redfield S, Austin JP, Akrami A. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. EClinicalMedicine. 2021. https://doi.org/10.1016/j.eclinm.2021.101019.
6. Al-Aly Z, Topol E. Solving the puzzle of Long Covid. Science. 2024;383(6685):830–2. https://doi.org/10.1126/science.adl0867.
7. Altmann DM, Whettlock EM, Liu S, Arachchillage DJ, Boyton RJ. The immunology of long COVID. Nat Rev Immunol. 2023;23(10):618–34.
8. Greenhalgh T, Sivan M, Perlowski A, Nikolich JŽ. Long COVID: a clinical update. The Lancet. 2024;404(10453):707–24.
9. Su S, Zhao Y, Zeng N, Liu X, Zheng Y, Sun J, Zhong Y, Wu S, Ni S, Gong Y, Zhang Z. Epidemiology, clinical presentation, pathophysiology, and management of long COVID: an update. Mol Psychiatry. 2023;28(10):4056–69.
10. Goldstein DS. Post-COVID dysautonomias: what we know and (mainly) what we don't know. Nat Rev Neurol. 2024;20(2):99–113.
11. Harrison PJ, Taquet M. Neuropsychiatric disorders following SARS-CoV-2 infection. Brain. 2023;146(6):2241–7.
12. Kubota T, Kuroda N, Sone D. Neuropsychiatric aspects of long COVID: a comprehensive review. Psychiatry Clin Neurosci. 2023;77(2):84–93.
13. Nugent K, Berdine G. Dyspnea and long COVID patients. Am J Med Sci. 2024. https://doi.org/10.1016/j.amjms.2024.07.024.
14. Komaroff AL, Lipkin WI. ME/CFS and Long COVID share similar symptoms and biological abnormalities: road map to the literature. Front Med. 2023;10:1187163.
15. Eaton-Fitch N, Rudd P, Er T, Hool L, Herrero L, Marshall-Gradisnik S. Immune exhaustion in ME/CFS and long COVID. JCI insight. 2024;9(20): e183810.
16. Spicer CM, Chu BX, Volberding PA, National Academies of Sciences, Engineering, and Medicine. Chronic Conditions Similar to Long COVID Long-Term Health Effects of COVID-19: Disability and Function Following SARS-CoV-2 Infection. Cambridge: National Academies Press (US); 2024.
17. Cantrell C, Reid C, Walker CS, Stallkamp Tidd SJ, Zhang R, Wilson R. Post-COVID postural orthostatic tachycardia syndrome (POTS): a new phenomenon. Front Neurol. 2024;15:1297964.
18. El-Rhermoul FZ, Fedorowski A, Eardley P, Taraborrelli P, Panagopoulos D, Sutton R, Lim PB, Dani M. Autoimmunity in long COVID and POTS. Oxford Open Immunol. 2023;4(1):002.
19. Goldenberg DL. How to understand the overlap of long COVID seminars in arthritis and rheumatism chronic fatigue syndrome/myalgic encephalomyelitis fibromyalgia and irritable bowel syndromes. Seminars Arthritis Rheumatism. 2024. https://doi.org/10.1016/j.semarthrit.2024.152455.
20. Perlis RH, Santillana M, Ognyanova K, Safarpour A, Trujillo KL, Simonson MD, Green J, Quintana A, Druckman J, Baum MA, Lazer D. Prevalence and correlates of long COVID symptoms among US adults. JAMA Netw Open. 2022;5(10):e2238804–e2238804.
21. Turk F, Sweetman J, Chew-Graham CA, Gabbay M, Shepherd J, van der Feltz-Cornelis C, STIMULATE-ICP Consortium. Accessing care for long Covid from the perspectives of patients and healthcare practitioners: a qualitative study. Health Expect. 2024;27(2):e14008.
22. O'Hare AM, Vig EK, Iwashyna TJ, Fox A, Taylor JS, Viglianti EM, Butler CR, Vranas KC, Helfand M, Tuepker A, Nugent SM. Complexity and challenges of the clinical diagnosis and management of long COVID. JAMA Netw Open. 2022;5(11):e2240332–e2240332.
23. Hamlin RE, Blish CA. Challenges and opportunities in long COVID research. Immunity. 2024;57(6):1195–214.
24. Ruß AK, Schreiber S, Lieb W, Vehreschild JJ, Heuschmann PU, Illig T, Appel KS, Vehreschild MJ, Krefting D, Reinke L, Viebke A. Genome-wide association study of post COVID-19 syndrome in a population-based study in Germany. Res Square. 2024;395:497.
25. Lammi, V., Nakanishi, T., Jones, S.E., Andrews, S.J., Karjalainen, J., Cortés, B., O'Brien, H.E., Fulton-Howard, B.E., Haapaniemi, H.H., Schmidt, A. and Mitchell, R.E., 2023. Genome-wide association study of long COVID. medRxiv, pp.2023–06.
26. Chaudhary, N.S., Weldon, C.H., Nandakumar, P., 23andMe Research Team, Holmes, M.V. and Aslibekyan, S., 2024.
27. Das S, Taylor K, Kozubek J, Sardell J, Gardner S. Genetic risk factors for ME/CFS identified using combinatorial analysis. J Transl Med. 2022;20(1):598.
28. Gardner S. Combinatorial analytics: an essential tool for the delivery of precision medicine and precision agriculture. Artif Intell Life Sci. 2021;1: 100003.
29. Das S, Taylor K, Beaulah S, Gardner S. Systematic indication extension for drugs using patient stratification insights generated by combinatorial analytics. Patterns. 2022;3(6):100456.
30. Taylor K, Pearson M, Das S, Sardell J, Chocian K, Gardner S. Genetic risk factors for severe and fatigue dominant long COVID and commonalities with ME/CFS identified by combinatorial analysis. J Transl Med. 2023;21(1):775.
31. Thompson RC, Simons NW, Wilkins L, Cheng E, Del Valle DM, Hoffman GE, Cervia C, Fennessy B, Mouskas K, Francoeur NJ, Johnson JS. Molecular states during acute COVID-19 reveal distinct etiologies of long-term sequelae. Nat Med. 2023;29(1):236–46.
32. Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, Jenkins G, Dishman E. The, "All of Us" research program. N Engl J Med. 2019;381(7):668–76.
33. The All of Us Research Program Genomics Investigators. Genomic data in the All of Us research program. Nature. 2024;627:340–6. https://doi.org/10.1038/s41586-023-06957-x.
34. Infinium Global Diversity Array-8 Kit Specifications. https://emea.illumina.com/products/by-type/microarray-kits/infinium-global-diversity.html last accessed 14 Jan 2025
35. Phillips R, Taiyari K, Torrens-Burton A, Cannings-John R, Williams D, Peddle S, Campbell S, Hughes K, Gillespie D, Sellars P, Pell B, Ashfield-Watt P, Akbari A, Seage CH, Perham N, Joseph-Williams N, Harrop E, Blaxland J, Wood F, Poortinga W, Wahl-Jorgensen K, James DH, Crone D, Thomas-Jones E, Hallingberg B. Cohort profile: the UK COVID-19 public experiences (COPE) prospective longitudinal mixed-methods study of health and well-being during the SARSCoV2 coronavirus pandemic. PLoS ONE. 2021;16(10): e0258484. https://doi.org/10.1371/journal.pone.0258484.
36. Blanchflower DG, Bryson A. Long COVID in the United States. PLoS ONE. 2023;18(11): e0292672. https://doi.org/10.1371/journal.pone.0292672.
37. Robertson MM, Qasmieh SA, Kulkarni SG, Teasdale CA, Jones HE, McNairy M, Borrell LN, Nash D. The epidemiology of long coronavirus disease in US adults. Clin Infect Dis. 2023;76(9):1636–45. https://doi.org/10.1093/cid/ciac961.

38. National Center for Health Statistics. U.S. Census Bureau, Household Pulse Survey, 2022–2024. Long COVID. Generated interactively 11 Jan 11, 2025. https://www.cdc.gov/nchs/covid19/pulse/long-covid.htm
39. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
40. Anderson C, Pettersson F, Clarke G, et al. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5:1564–73. https://doi.org/10.1038/nprot.2010.116.
41. Grinde KE, Browning BL, Reiner AP, Thornton TA, Browning SR. Adjusting for principal components can induce collider bias in genome-wide association studies. PLoS Genet. 2024;20(12): e1011242. https://doi.org/10.1371/journal.pgen.1011242.
42. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. Statist Med. 2015;34(28):3769–92. https://doi.org/10.1002/sim.6605.
43. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. J R Stat Soc B. 1995;57:289–300.
44. Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika. 1928;20A:175–240.
45. Fontes-Dantas FL, Fernandes GG, Gutman EG, De Lima EV, Antonio LS, Hammerle MB, Mota-Araujo HP, Colodeti LC, Araújo SMB, Froz GM, da Silva TN, Duarte LA, Salvio AL, Pires KL, Leon LAA, Vasconcelos CCF, Romão L, Savio LEB, Silva JL, da Costa R, Clarke JR, Da Poian AT, Alves-Leon SV, Passos GF, Figueiredo CP. SARS-CoV-2 Spike protein induces TLR4-mediated long-term cognitive dysfunction recapitulating post-COVID-19 syndrome in mice. Cell Rep. 2023;42(3): 112189. https://doi.org/10.1016/j.celrep.2023.112189.
46. Mukherjee S. Toll-like receptor 4 in COVID-19: friend or foe? Future Virol. 2022. https://doi.org/10.2217/fvl-2021-0249.
47. Liu ZM, Yang MH, Yu K, Lian ZX, Deng SL. Toll-like receptor (TLRs) agonists and antagonists for COVID-19 treatments. Front Pharmacol. 2022;7(13): 989664. https://doi.org/10.3389/fphar.2022.989664.
48. Ustinova M, Peculis R, Rescenko R, Rovite V, Zaharenko L, Elbere I, Silamikele L, Konrade I, Sokolovska J, Pirags V, Klovins J. Novel susceptibility loci identified in a genome-wide association study of type 2 diabetes complications in population of Latvia. BMC Med Genom. 2021;14(1):18. https://doi.org/10.1186/s12920-020-00860-4.
49. de Goede KE, Harber KJ, Gorki FS, Verberk SGS, Groh LA, Keuning ED, Struys EA, van Weeghel M, Haschemi A, de Winther MPJ, van Dierendonck XAMH, Van den Bossche J. d-2-Hydroxyglutarate is an anti-inflammatory immunometabolite that accumulates in macrophages after TLR4 activation. Biochim Biophys Acta Mol Basis Dis. 2022;1868(9): 166427. https://doi.org/10.1016/j.bbadis.2022.166427.
50. Sardell J, Das S, Taylor K, Stubberfield C, Malinowski A, Strivens M, Gardner S. Actively protective combinatorial analysis: a scalable novel method for detecting variants that contribute to reduced disease prevalence in high-risk individuals. Artif Intell Life Sci. 2025;7: 100125. https://doi.org/10.1016/j.ailsci.2025.100125.
51. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20(8):467–84. https://doi.org/10.1038/s41576-019-0127-1.
52. Moreno-Grau S, Vernekar M, Lopez-Pineda A, Mas-Montserrat D, Barrabés M, Quinto-Cortés CD, Moatamed B, Lee MTM, Yu Z, Numakura K, Matsuda Y, Wall JD, Ioannidis AG, Katsanis N, Takano T, Bustamante CD. Polygenic risk score portability for common diseases across genetically diverse populations. Hum Genomics. 2024;18(1):93. https://doi.org/10.1186/s40246-024-00664-y.
53. Clifton L, Collister JA, Liu X, et al. Assessing agreement between different polygenic risk scores in the UK biobank. Sci Rep. 2022;12:12812. https://doi.org/10.1038/s41598-022-17012-6.
54. Curtis D. Clinical relevance of genome-wide polygenic score may be less than claimed. Ann Hum Genet. 2019;83(4):274–7. https://doi.org/10.1111/ahg.12302.
55. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015;7:16. https://doi.org/10.1186/s13073-015-0138-2.
56. Aliferis C, Simon G. Overfitting underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In: Simon GJ, Aliferis C, editors. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences. Cham: Springer; 2024.
57. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. Biometrics. 2022;78(1):214–26. https://doi.org/10.1111/biom.13400.
58. Burstein D, Hoffman G, Mathur D, Venkatesh S, Therrien K, Fanous AH, Bigdeli TB, Harvey PD, Roussos P, Voloudakis G. Detecting and adjusting for hidden biases due to phenotype misclassification in genome-wide association studies. medrxiv. 2023. https://doi.org/10.1101/2023.01.17.23284670.
59. All of Us Genomic Quality Report, Feb 3. 2025. https://support.researchallofus.org/hc/en-us/articles/29390274413716-All-of-Us-Genomic-Quality-Report last Accessed 12 Apr 2025
60. Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, Linsell L, Staplin N, Brightling C, Ustianowski A, Elmahi E, Prudon B, Green C, Felton T, Chadwick D, Rege K, Fegan C, Chappell LC, Faust SN, Jaki T, Jeffery K, Montgomery A, Rowan K, Juszczak E, Baillie JK, Haynes R, Landray MJ. Dexamethasone in hospitalized patients with Covid-19. N Engl J Med. 2021;384(8):693–704. https://doi.org/10.1056/NEJMoa2021436.
61. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12:44. https://doi.org/10.1186/s13073-020-00742-5.

Publisher's Note