nature research

Peer Review File

Manuscript Title: Genetic mechanisms of critical illness in Covid-19

Editorial Notes:

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

Summary: The authors describe a genome-wide association study of host genetic factors contributing to critical illness in COVID-19 patients in a multi-ethnic (although predominantly European) cohort of 2,224 critically ill patients compared to UK biobank controls at a ratio of 1:5. This analysis identifies 15 independent loci associated with critical illness, 8 of which are validated through sensitivity analyses using independent population cohorts as controls (100,000 Genomes and Generation Scotland). The authors then provide further replication for three new loci (OAS cluster, DPP9, IFNAR2) by combining their associations results with those of the COVID-19 HGI. Meta-analysis of the GenOMICC European sample with COVID-HGI and 23 & Me data validate the TYK2 locus on chr 19 and uncover associations in the MHC region and NMNAT3. Mendelian randomization analysis interrogating overlap of GWAS associations with genetic regulation of gene expression from GTEx suggests a role for expression of IFNAR2 and speculative evidence for TYK2. In a transcriptome-wide associations analysis, the authors show further support for expression of 18 genes including several reported to be involved in SARS-CoV2 replication and response in other studies. Finally, the authors show significant heritability of critical COVID-19 illness and potential genetic overlap with known cofactors such as adiposity and educational attainment.

Overall this is a well-run study of obvious impact and importance and I am very pleased to see that the results have already been shared with the COVID-19 HGI. I do however have several comments I feel will improve confidence in the associations analyses and overall clarity of the work.

Major comments
Some further description of how cases and controls were matched is warranted. In the methods, the authors state using controls from the UK biobank with the same ancestry designations as cases. Given the potential diversity within these broad population labels, wouldn't a matching strategy based on PC clustering be more rigorous? My concerns here could also be addressed by presentation of ancestry-specific case/control PCA plots that demonstrate good clustering. Also, were age and sex included in the matching strategy? Apologies if this details was specified and I missed it.

Similarly, I would like to see more attention paid to the quality of the genome-wide association results. Specifically in figure 1 the QQ plot seems to show evidence for some inflation in the overall test statistic with departure from the null expectation obvious around $p<10\text{-}2$. As well, there are loci in the Manhattan plot, notably on chromosome 12, that pass the statistical threshold for significance but appear to be spurious, and do not replicate in follow-up analyses suggesting. This is even more pronounced in the multi-ethic analysis presented in figure S1 where variants on chr5, 7, 11, 12, 14, 15 and X pass the specified statistical threshold. These results, if assumed to be true, warrant discussion, or, if assumed to be false, suggest some additional confounding. I would like to see QQ plots for all separate ancestry GWAS used in the meta analysis, as well as genomic inflation factors, and results from linear mixed model analysis to ensure against broad inflation.

I'm not tremendously enthusiastic about the claims made in the transcriptome-wide Mendelian randomization analysis. As the author themselves point out, no genes pass the threshold for statistical significance given the number of tests run and the evidence for TYK2 relies on re-using the same GTEx expression set in both analyses. I suggest moving this result fully to the supplement. Also in this section, I'm curious why the authors

used a curated set of target genes relating to host-targeted drugs rather than interrogating the evidence for genetic regulation of expression in the gene candidates identified through GWAS. It seems to me this would be a higher priority analysis.

Finally, given that in the introduction the authors suggest that the phenotype under study here is distinct from mild or other severe COVID-19 related illness, I wonder if analysis of overlap of results between their data, COVID-19 HGI and 23&Me would provide additional isight into this claim. I.e. are their loci that can be confidently attributed to their critical illness phenotype beyond those observed for hospitalization (COVID-HGI) or the 23&Me broad respiratory phenotype or are all three analogous with the GenOMICC phenotype providing more power?

Minor
In the introduction, the language surrounding the precise phenotype is a bit vague. Specifically, lines 114-115 "…we performed a genome-wide association study comparing to controls from population genetic studies in the UK". A clearer statement on the phenotype here would enhance readability.

It is not always clear in the main text when the authors are presenting analyses in European ancestry only individuals (replication with additional control groups, sex-specific analysis). This should be explicitly stated.

Sample number reported in the main text may lead to confusion. The authors report the total number of participants genotyped, 2734, but not the number after QC (2,244) which is listed in the abstract and in the Table 1 header. As well the actual number of controls used per ancestry group and in total is not presented.

I found the sex-specific analysis paragraph initially confusing. as written, it seems to report that LZTFL1, CCHCR1 and NOTHC4 show sex specificity but the data in Table S1 show very consistent effects across sex when accounting for differential power due to sample size. Unless I'm mistaken the data don't show strong evidence for sex-specific effects. This should be clarified.

I would like to see consistency in reporting of the effect sizes. The authors switch back and for the between presenting OR (Table 2) beta (Table 3) and both (Table 4). My preference in a binary association study would be OR plus 95% CI in all main text tables. I believe this is the most widely understood effect estimate for a general audience.

Table 2, why are the Generation Scotland results not presented? Seems important for completeness.

Table 3 legend mentions 23 & me data that are not presented.

Figures 2 and 3 could be improved for clarity. Gene name locations often overlap with other display items.

The MR expression section should state the GTEx tissues used in the analysis.

The link to the summary results is not active.

Referee #2 (Remarks to the Author):

The manuscript presents the largest GWAS to date on critical illness upon COVID-19 infection. The data has been carefully collected and the "recycled" UKB control samples while not being ideal (see comments below), allowed faster results and seemed to have been adequately corrected for. QC procedures and imputation have been done with state-of-the-art methods, so was the GWAS performed. The post-GWAS analyses are mostly thorough and open up interesting avenues for medical interventions, but of course this is only the first step in a very long process. Below I list my comments which could help improve the manuscript.
Major comments:

With such modern imputation panels (such as TopMed), the significance threshold should be adjusted to more realistic 1E-8, not the old 5E-8 (which was created for older arrays and poorer imputation panels).

Any explanation for the non-replication (and significantly different OR) of half of the loci [Table 3]? Can it be due to the non-ideal design of all cases genotyped on an Illumina array, imputed with TopMed, while controls are on an Affy array, imputed using a mix UK10K, 1000G panel? A more informative Table (than Table 2) would be useful in the Supplement, including case RAF, UKBB control RAF, 100000 genomes RAF, GS RAF, 1KG RAF, etc.

The expression MR analysis is poorly described: which tissues were considered? Which MR method was used? (It is only mentioned in the Methods, but not in the Results – same for metaXcan for TWAS) What is meant by "were then tested for independent external evidence", which then turns out to be only partially independent (because exposure effects were taken from the same GTEx (blood? lung?) sample? Why not the meta-analysed outcome effect sizes were used? The fact that the IFNAR2 Heidi P-value is 0.015, only three times higher than the SMR P-value, indicates that the signal may be driven by a single outlier SNP. Could the authors show the classical beta_exposure vs beta_outcome plot for those 6 SNPs? Which drug is targeting IFNAR2? Also, have the authors applied any more robust MR method (e.g. median/mode-based estimators)?

Wouldn't protein-> severity be a better MR exercise? There are many studies with available protein QTL summary stats. Also, using such a small sample as GTEx is prohibitive for statistical power. Have the authors considered using the summary stats of the eQTL-Gen consortium (whole blood, n=32K)? Also, why only gene expression was explored as exposure: how about other potential modifiable risk factors (diabetes, FEV, obesity, etc.)?

It is somewhat surprising how MR showed nothing GW significant, but "combined meta-TWAS analysis" revealed 18 genes. Mathematically TWAS and MR (standard IVW without any heterogeneity filtering) are equivalent. Is it the combination of blood and lung tissue that made the difference? Figure 3 confirms my confusion: how come the lung-only TWAS shows much weaker signals than the meta-TWAS. It would mean that these are mostly driven by blood-TWAS? Can I see the QQ-plot of the meta-TWAS P-values?

The MAIC analysis remains extremely vague: a single (hard-to-read) plot is provided in the supplement, but no exact quantification of statistical evidence of overrepresentation is given, neither significance (or confidence interval) attached to it.

The fact that education and intelligence show the strongest genetic correlation worries me slightly, as it is typically an indicator of uncorrected population stratification and mostly proxies of socio-economic status. Also, how come overall health rating has positive genetic correlation with severe complications?

The tissue enrichments look all over the place, I do not see any clear pattern, and barely anything emerges to be significant. Moreover, I do not see the link with spleen/pancreas and COVID19.

I have the impression that only the EUR subset was used in the entire manuscript, have I missed something? If not, why were the other cases ignored?

I do not see any pathway enrichment analysis in the paper. It guess it has been attempted, but nothing showed up?

Given the 1667 sequenced case, could the author compare the frequency of rare variants to those in gnomAD?
Minor comments:

1. It has to be explicitly stated what is meant by "equivocal evidence of heterogeneity" for the Heidi test – what heterogeneity P-value they got, have they removed any instruments as a consequence?
2. Regarding SNP-based heritability: please specific on which scale was it estimated. It is never mentioned in the HDL paper either.
3. The authors repeatedly use "discovered and replicated", but it is never clearly mentioned how many loci do they claim as new discovery and which ones are replicating previous ones.
4. "summary statistics 221 from GenOMICC are openly available": A Link would be great there. As it is not on the GenOMICC website.

5. Why replication P-values and betas are not listed for 23andMe in Table 3?
6. Figure 3 would be more informative to show gcc on top and hgi+23andMe on the bottom in the Miami plot.

Referee #3 (Remarks to the Author):

This manuscript by Baillie and colleagues identifies novel loci and confirms loci associated with critical illness from COVID-19. Using population based controls the authors evaluate the host genetic and transcriptomic associations with severe COVID-19 outcomes requiring intensive care follow up from across more than 200 clinical hospital intensive care sites in the UK.

The authors should be credited for acknowledging the heterogeneity in COVID-19 outcomes, and for focusing on one extreme--hospitalized cases. By using an established network of collaboration (and expertise) for critical illness across the UK, they are able to identify a uniform definition of cases.

* However, its not clear from the data/table provided that this population of cases is homogeneous. Additional details on the need for mechanical ventilation, or oxygen support, or days (with range) of hospitalization would be helpful to understand if all of these patients are similar, or still represent a spectrum of hospitalized outcomes.

* Similarly, the time frame of recruitment into the study in light of the epidemic in the UK would be helpful. Because, we know globally that those affected prior to governments shutting businesses etc or the adoption of social distancing are different than those later (i.e age), and treatment also evolved with more provider experience.

The one area that the authors address but do not provide details on, are the co-morbidities. We know these are huge risk factors for severe outcomes along with age and sex.
* Although the authors indicate in Table 1 the "significant comorbidities" its not clear what that means? In studies out of China and the United States--hypertension is a risk factor. Is that considered in the significant comorbidities? For example, were the majority of comorbidities COPD or Asthma--in which case some of these loci really represent a novel loci for severe hospitalized COPD exacerbation following a viral infection? More details here will also educate us on who was sickest and if genetics goes beyond these comorbidities.

Baillie and colleagues identify the data rich UK Biobank as a source of population based controls. They select 5 controls per case.
* However, details on these "controls" with regards to age, comorbidities and sex should be provided in a table. If they were younger and healthier it does suggest that they would have been unlikely anyway to be in the intensive care units at least in the early days of the pandemic. And what about obesity? Was this matched on?

* Interestingly, the authors have access to COVID-19 positive individuals in the UK Biobank, but chose to exclude them. What was the rationale? This group of individuals would be the best match, and ideally should be considered since they were exposed, and did not require hospitalization. If we are trying to identify a severity locus--this would be the best approach.

* Using population based controls in the manner of 5:1 may increase sample size but the authors should acknowledge the underlying assumption is that they believe that these population controls may be exposed but not require hospitalization. Otherwise there is inherent misclassification of the controls.

*The authors also only present in the main figures the European plots and locus zooms (not all locus zooms look convincing). Given the worldwide pandemic that affects all ancestry groups, its seems crucial that the genetics for all groups is also presented in the main portion of this paper. It is a strength of this study. And it could also aid our understanding of the Chromosome 3 locus--if credible sets were used to narrow down this region. This could be done (or attempted) with this diverse population and aid the overall genetic community.
*The PCA plots supplementary should also show each ancestry separately with controls so that we can see the overall distributions, and more than just PC1 and PC3.

The addition of the transcriptomics and gene grouping is interesting and seems adequate in approach. The heritability analysis is likely lacking sufficient power and that should be noted as a limitation and not reflective of a lack of heritability given the effect sizes and that this is SNP based.

But the heritability correlation is very intriguing.
* The authors cite correlations with education and intelligence and with body mass measurements. These alone are reasons why the cases/controls should be matched on these factors, or ideally a propensity score matching approach should be done. We know that different people were exposed at different times based on their occupation, public transportation, etc and we need to address those differences. Similarly, we know BMI is a risk factor for severe outcomes, so this should be addressed in the analysis. The authors do excellent work to identify these measures but now must be sure that they are not influencing or confounding their genetic association also.

Overall, this is an easy paper to read and is clear in writing style. Many different genomics approaches were employed, but the overall message is that new loci have been identified. To solidify those loci, additional epidemiologic and clinical descriptions and inclusion in the analysis are necessary.

Minor note: some of the references are not correctly annotated (or there is an issue with how they have printed--with extra characters in the citation).

Author Rebuttals to Initial Comments:

1 Referee #1

1.1 Summary

The authors describe a genome-wide association study of host genetic factors contributing to critical illness in COVID-19 patients in a multi-ethnic (although predominantly European) cohort of 2,224 critically ill patients compared to UK biobank controls at a ratio of 1:5. This analysis identifies 15 independent loci associated with critical illness, 8 of which are validated through sensitivity analyses using independent population cohorts as controls (100,000 Genomes and Generation Scotland). The authors then provide further replication for three new loci (OAS cluster, DPP9, IFNAR2) by combining their associations results with those of the COVID- 19 HGI. Meta-analysis of the GenOMICC European sample with COVID-HGI and 23 & Me data validate the TYK2 locus on chr 19 and uncover associations in the MHC region and NMNAT3. Mendelian randomization analysis interrogating overlap of GWAS associations with genetic regulation of gene expression from GTEx suggests a role for expression of IFNAR2 and speculative evidence for TYK2. In a transcriptome-wide associations analysis, the authors show further support for expression of 18 genes including several reported to be involved in SARS-CoV2 replication and response in other studies. Finally, the authors show significant heritability of critical COVID-19 illness and potential genetic overlap with known cofactors such as adiposity and educational attainment.

Overall this is a well-run study of obvious impact and importance and I am very pleased to see that the results have already been shared with the COVID-19 HGI. I do however have several comments I feel will improve confidence in the associations analyses and overall clarity of the work.

1.2 Major comments

1.2.1 PCA plots

Some further description of how cases and controls were matched is warranted. In the methods, the authors state using controls from the UK biobank with the same ancestry designations as cases. Given the potential diversity within these broad population labels, wouldn't a matching strategy based on PC clustering be more rigorous? My concerns here could also be addressed by presentation of ancestry-specific case/control PCA plots that demonstrate good clustering.

We apologise for not being clearer in our description. We did use PCA to determine genetic ancestry, and then assigned labels to PCA clusters. We have amended the description of these groupings to make this clearer. We have also provided additional PCA plots for each ancestry group as suggested, in Supplementary Figure 3.

We performed a first PCA using the 1000G phase 3 samples as sole input for the axis of genetic variation to infer (by projection) ancestries of cases. The cases were then matched to UKBB participants of the same broad ancestry groups. The first ten derived PC were used as covariates in the GWAS analyses.

The new PCA plots are reproduced here for convenience: Figure 1.

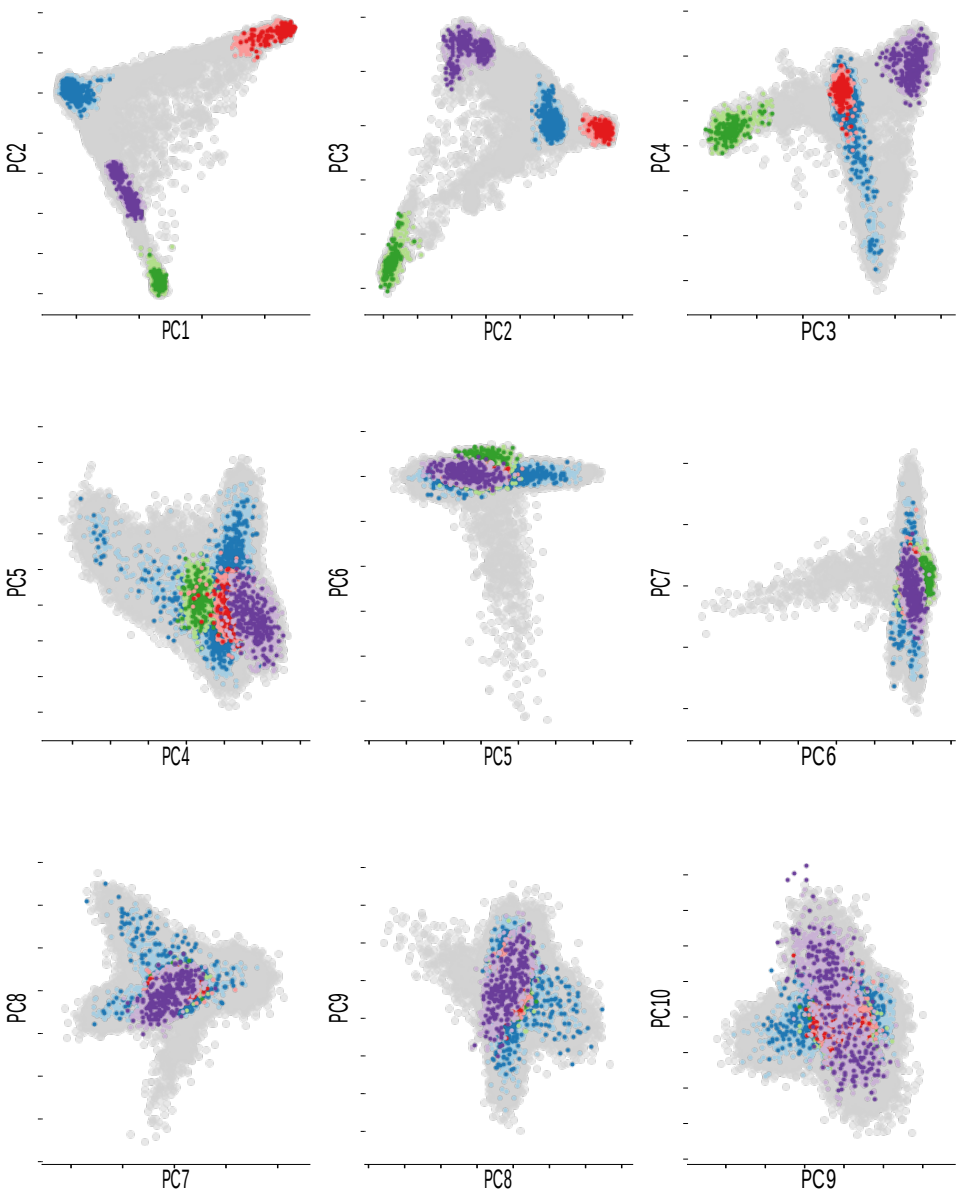


Figure 1: PCA plots showing the distribution of all cases and controls for the first 10 principal components. Cases are shown as coloured closed circles: European (EUR, blue), African (AFR, red), East Asian (EAS, green), and South Asian (SAS, purple). Controls for each ancestry group are shown as closed circles in a lighter shade of the colour for that ancestry group. UK Biobank population background is shown as light grey closed circles.

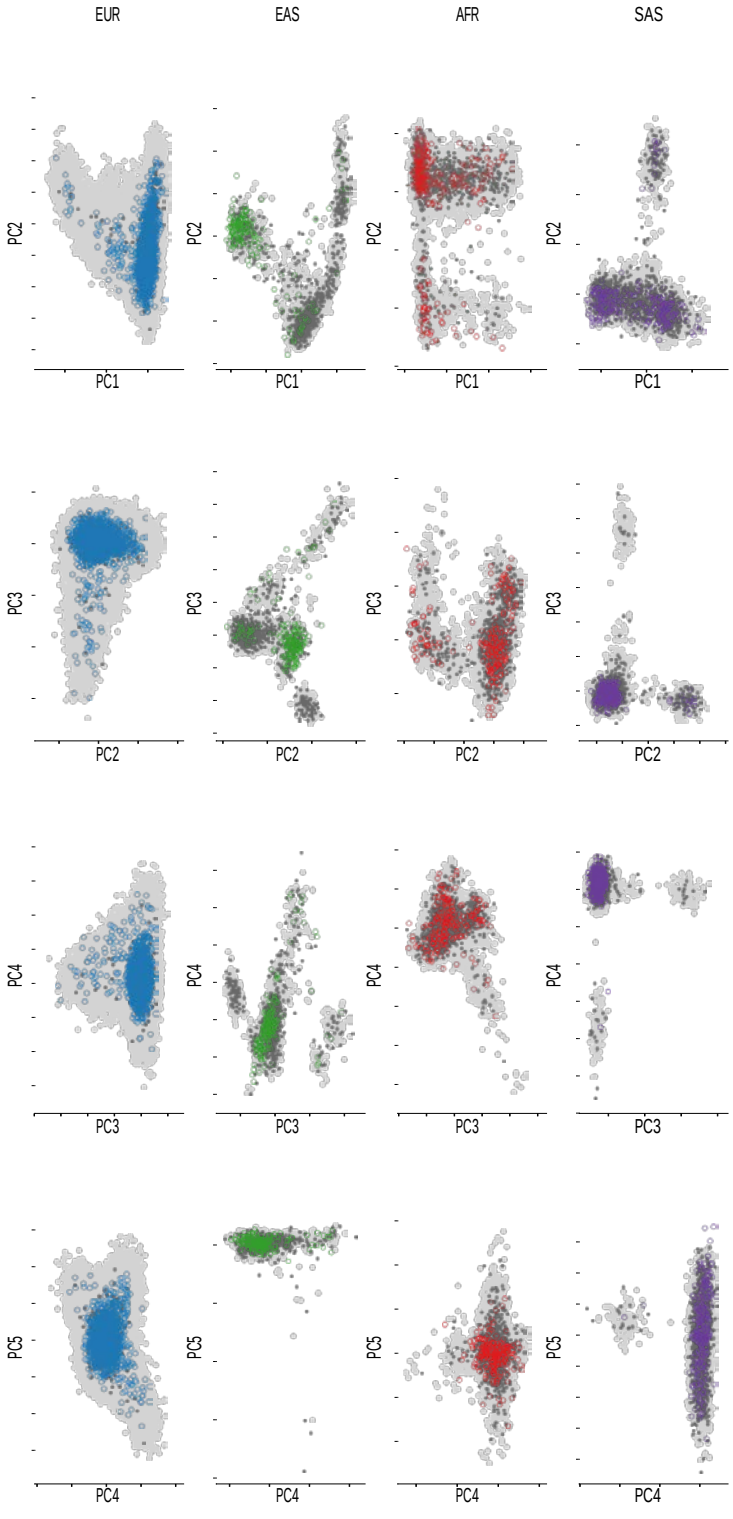


Figure 2: PCA plots showing the distribution of all cases and controls for the first 5 principal com- ponents for each ancestry group. Cases are shown as coloured open circles: European (EUR, blue), African (AFR, red), East Asian (EAS, green), and South Asian (SAS, purple). Controls are dark grey closed circles. UK Biobank population background is shown as light grey closed circles.

> Also, were age and sex included in the matching strategy? Apologies if this details was specified and I missed it.

Since the age range of UK Biobank participants is narrow, and different from the GenOMICC cases, age matching would result in the loss of a substantial number of cases (see sec. 3.2.2). We therefore did not use age for matching in the primary analysis but instead used age, mean-centered age^2, and sex as covariates in the logistic regression model.

However we have also now repeated the GWAS using a stringent matching strategy in response to feedback from reviewers: please see sec. 3.2.2 for a description of the approach and results.

1.2.2 Replication

> Similarly, I would like to see more attention paid to the quality of the genome-wide association results. Specifically in figure 1 the QQ plot seems to show evidence for some inflation in the overall test statistic with departure from the null expectation obvious around p<10-2. As well, there are loci in the Manhattan plot, notably on chromosome 12, that pass the statistical threshold for significance but appear to be spurious, and do not replicate in follow-up analyses suggesting. This is even more pronounced in the multi-ethic analysis presented in figure S1 where variants on chr5, 7, 11, 12, 14, 15 and X pass the specified statistical threshold. These results, if assumed to be true, warrant discussion, or, if assumed to be false, suggest some additional confounding. I would like to see QQ plots for all separate ancestry GWAS used in the meta analysis, as well as genomic inflation factors, and results from linear mixed model analysis to ensure against broad inflation.

We have included QQ plots and genomic inflation factors for all ethnic groups in Supplementary Figure 5. These are reproduced here for convenience: Figure 3.

In two ancestry groups, both of which have < 200 cases so far in GenOMICC (EAS and SAS), there is substantial inflation, which likely reflects confounding arising from the systematic differences between cases and the opportunistic controls included in our study (see new PCA plots in Supplementary Figure 3, reproduced in Figure 1). There is also some residual inflation in the primary analysis in EUR with $\lambda 0.5 = 1.099$ (Figure 2), which we discuss in more detail below (sec. 2.2.5).

In order to mitigate the effect of this on the new discoveries reported in our manuscript, we have:

- restricted the initial discovery set to the largest and best-matched ancestry group (EUR);
- confirmed the signals are present, and in the same direction, in three distinct control groups;
- replicated our analyses in the largest and most-relevant datasets available: Covid19 HGI and 23andMe;
- performed a full meta-analysis combining GenOMICC, Covid19 HGI and 23andMe which confirms the GenOMICC hits with stronger associations.

However we agree that it is important to include the analysis of multiple ethnic groups where possible, both in order to do the best we can to maximise the relevance of this work to all humans, and also because of our ethical obligation to the patients who wished to contribute to this research. These unconfirmed results may have value for future investigations into host genetics in Covid-19 in ancestry groups that are not well-covered by GenOMICC in the UK.

Five out of 8 variants in the replication table are successfully replicated (Table 3). It should be noted that more genome-wide significant variants appear on the Manhattan plot (Figure 1) but did not have confirmatory signals in the comparisons with additional control groups (Generation Scotland

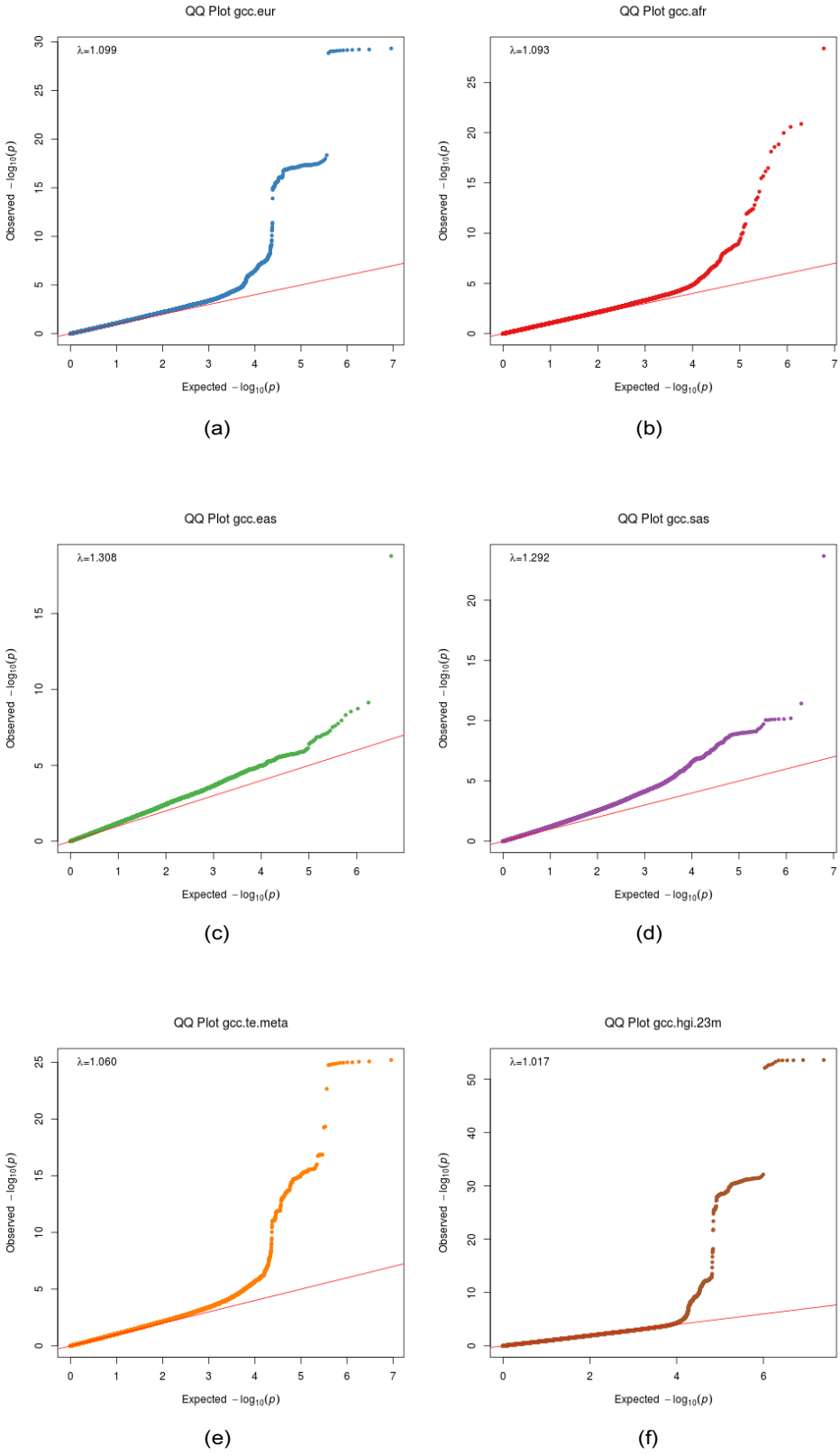


Figure 3: Q:Q plots for each ancestry group in GenOMICC: gcc.eur - European; gcc.afr - African; gcc.eas - East Asian; gcc.sas - South Asian, together with trans-ethnic meta-analysis (gcc.te.meta), and meta-analysis comprising GenOMICC, HGI and 23andMe data (gcc.hgi.23m). - genomic inflation value.

and 100,000 genomes). We decided to include these variants on the Manhattan plot to give a clear indication of effect of this validationn step.

Three out of these 8 variants do not replicate. All three are on chromosome 6, in a region of the genome known to be particularly senstive to population structure (and important in infectious disease). One of these, in *CCHCR1*, tends towards significance in the replication analysis, and is biologically plausible. It may be that this variant fails to replicate due to a combination of the winner's curse, and to the very small numbers of critically ill patients recruited to the replication studies. Only 2 studies contribute critically ill cases to the Covid-19 HGI - GENCOVID_EUR: 327 cases, 2461 controls; and BRACOVID_AMR: 101 cases, 1533 controls. Combined, these studies do not provide any suggestion of genetic association signals. The other variants, one in the MHC and one in *NOTCH4*, may also be subject to these limitations, or they may be spurious associations arising, for example, due to genotyping errors or population structure. We have discussed these problems further in sec. 3.2.2.

We have substantially mitigated against genotyping errors by using three different control groups genotyped on different platforms (UK Biobank, Generation Scotland and 100,000 genomes; Table 2). This also protects against the effect of population structure affecting the genome-wide significant associations.

> I'm not tremendously enthusiastic about the claims made in the transcriptome-wide Mendelian randomization analysis. As the author themselves point out, no genes pass the threshold for statistical significance given the number of tests run and the evidence for TYK2 relies on re- using the same GTEx expression set in both analyses. I suggest moving this result fully to the supplement.

We agree that the partial replication of this result, using a new GWAS dataset, but the same expression data, was not ideal. We have now repeated the analysis using a different expression dataset to provide a full replication, which robustly confirms these associations (please see further discussion below, sec. 2.2.3).

> Also in this section, I'm curious why the authors used a curated set of target genes relating to host-targeted drugs rather than interrogating the evidence for genetic regulation of expression in the gene candidates identified through GWAS. It seems to me this would be a higher priority analysis.

In normal circumstances we agree we might prefer to take the approach suggested. However, during the current outbreak, our first objective is to provide evidence that will help to prioritise treatments for use in clincal trials. Several of the GenOMICC investigators are also investigators on the RECOVERY trial,[1,2] and are faced with the challenge of choosing potential treatments targeting the host response, with very little evidence to support each one. For this reason we chose to focus on a short candidate list of directly therapeutically-informative genes.

> Finally, given that in the introduction the authors suggest that the phenotype under study here is distinct from mild or other severe COVID-19 related illness, I wonder if analysis of overlap of results between their data, COVID-19 HGI and 23&Me would provide additional isight into this claim. I.e. are their loci that can be confidently attributed to their critical illness phenotype beyond those observed for hospitalization (COVID-HGI) or the 23&Me broad respiratory phenotype or are all three analogous with the GenOMICC phenotype providing more power?

These trends are interesting and potentially important, so we have provided details in Supplementary

Figure 6 to allow readers to explore these possibilities. There are several important caveats. Firstly, the critical illness phenotype is a small subset of the hospitalised cohorts in each of the other studies (HGI and 23andMe), so with sufficient numbers, eventually, these same associations may be discovered in these studies. Where the signals are less strong among hospitalised cases, there is no way to distinguish a continuous effect (in which enrichment for risk alleles increases with each step in worsening respiratory failure) from a dichotomous one (in which patients with respiratory failure have completely distinct susceptibility factors) with the existing data.

1.3 Minor comments

In the introduction, the language surrounding the precise phenotype is a bit vague. Specifically, lines 114-115 "…we performed a genome-wide association study comparing to controls from population genetic studies in the UK". A clearer statement on the phenotype here would enhance readability.

We have altered the text as suggested.

It is not always clear in the main text when the authors are presenting analyses in European ancestry only individuals (replication with additional control groups, sex-specific analysis). This should be explicitly stated.

We have clarified this in the manuscript and included this sentence in the results section: "The largest ancestry group contained 1676 individuals of European descent (EUR); this group were used for the primary analyses presented below."

Sample number reported in the main text may lead to confusion. The authors report the total number of participants genotyped, 2734, but not the number after QC (2,244) which is listed in the abstract and in the Table 1 header. As well the actual number of controls used per ancestry group and in total is not presented.

We have clarified this in the second paragraph of the results section.

I found the sex-specific analysis paragraph initially confusing. as written, it seems to report that LZTFL1, CCHCR1 and NOTHC4 show sex specificity but the data in Table S1 show very consistent effects across sex when accounting for differential power due to sample size. Unless I'm mistaken the data don't show strong evidence for sex-specific effects. This should be clarified.

Yes, we agree with the reviewer's interpretation that our study did not detect sex-specific effects. We have simplified the main manuscript to say this clearly: "A sex-specific GWAS among this group found no sex-specific associations".

I would like to see consistency in reporting of the effect sizes. The authors switch back and for the between presenting OR (Table 2) beta (Table 3) and both (Table 4). My preference in a binary association study would be OR plus 95% CI in all main text tables. I believe this is the most widely understood effect estimate for a general audience.

We have replaced these values in the results tables as suggested.

> Table 2, why are the Generation Scotland results not presented? Seems important for complete- ness.

We have added the Generation Scotland p-values as requested.

> Table 3 legend mentions 23 & me data that are not presented.

We have now corrected this to include a meta-analysis of HGI and 23andMe.

> Figures 2 and 3 could be improved for clarity. Gene name locations often overlap with other display items.

We have re-created all manhattan and z-score plots with more visible gene names.

> The MR expression section should state the GTEx tissues used in the analysis.

We apologise for not being clearer about this in the results section - Reviewer 2 raised the same issue and we have responded in full in sec. 2.2.3.

> The link to the summary results is not active.

Many thanks for drawing this to our attention. We have already shared the results directly with investigators around the world, and have corrected the website so that the link at https://genomicc.org/data is now active.

2 Referee #2

2.1 Summary

> The manuscript presents the largest GWAS to date on critical illness upon COVID-19 infection. The data has been carefully collected and the "recycled" UKB control samples while not being ideal (see comments below), allowed faster results and seemed to have been adequately corrected for. QC procedures and imputation have been done with state-of-the-art methods, so was the GWAS performed. The post-GWAS analyses are mostly thorough and open up interesting avenues for medical interventions, but of course this is only the first step in a very long process. Below I list my comments which could help improve the manuscript.

2.2 Major comments

2.2.1 P-value threshold

> With such modern imputation panels (such as TopMed), the significance threshold should be adjusted to more realistic 1E-8, not the old 5E-8 (which was created for older arrays and poorer imputation panels).

As suggested, we have now used the lower threshold of 1×10^{-8} as the final test for significance in the meta-analysis results reported in Table 4 and Figure 2. Since we have only considered alleles with minor allele frequency $> 1\%$, we have used the standard sigificance threshold of 5×10^{-8} for the decision to include variants in the replication test. In this analysis we only consider common variants, regardless of the imputation panel used. The validity of this approach is supported by the replication in external datasets and the strong signals in meta-analysis at these loci.

2.2.2 Replication

> Any explanation for the non-replication (and significantly different OR) of half of the loci [Table 3]? Can it be due to the non-ideal design of all cases genotyped on an Illumina array, imputed with TopMed, while controls are on an Affy array, imputed using a mix UK10K, 1000G panel?

Reviewer 1 made a similar point, so we discuss this in more detail above: sec. 1.2.2. In summary we agree that the pragmatic design of this study is not ideal, but we can be very confident that the replicated findings are real: the use of two distinct control groups mitigates the specific risks mentioned by the reviewer, and there is a real possibility that the relative paucity of extreme-susceptibility cases from the replication cohorts is the fundamental limitation in our replication efforts. Nonetheless, 5 out of the 8 new associations from GenOMICC (and all of the variants outside of the major histocompatibility complex) have robustly replicated in external studies.

> A more informative Table (than Table 2) would be useful in the Supplement, including case RAF, UKBB control RAF, 100000 genomes RAF, GS RAF, 1KG RAF, etc.

We have added a downloadable csv file reporting the full details as requested.

2.2.3 Mendelian randomisation

> The expression MR analysis is poorly described: which tissues were considered? Which MR method was used? (It is only mentioned in the Methods, but not in the Results – same for metaXcan for TWAS)

We apologise that this was not made clear in the results section and have now added text to clarify this. In both cases we did state this in the Materials and Methods section:

- GTEx (v7) whole blood expression results were used as the exposure.
- Two-sample summary data based Mendelian randomisation was performed as described in detail by Zhu *et al*.[3]

> What is meant by "were then tested for independent external evidence", which then turns out to be only partially independent (because exposure effects were taken from the same GTEx (blood? lung?) sample?

We have now undertaken a full replication of our SMR findings using the results of eQTL-gen (with many thanks to the reviewer for this suggestion). The MR associations with IFNAR2 and TYK2 fully replicate. The results are now described in the main paper:

- *IFNAR2*: replication p-value 7.46E-04
- *TYK2*: replication p-value 5.53E-05

Why not the meta-analysed outcome effect sizes were used?

This was a pragmatic decision in order to mitigate against falsely inferring heterogeneity of effect-size estimates due to LD differences between the tested populations: GTEx is predominantly drawn from individuals of European ancestry,[4] and our LD reference entirely so.

2.2.4 HEIDI test for heterogeneity

The fact that the IFNAR2 Heidi P-value is 0.015, only three times higher than the SMR P-value, indicates that the signal may be driven by a single outlier SNP.

We interpret a significant HEIDI statistic as indicating is that there is heterogeneity in the SMR effect-size estimates between the lead SNP and at least one of the other SNPs included in the HEIDI test, not necessarily that the lead-SNP is an outlier.

In order to make this clear to readers we have added the following caveat: "There was equivocal evidence of heterogeneity (HEIDI[3] $p = 0.015$), indicating that the effect of this variant on critical illness in Covid-19 may be mediated through another mechanism, which may lead to an under- or over-estimation of the effect of IFNAR2 expression on risk of critical illness."

Could the authors show the classical beta_exposure vs beta_outcome plot for those 6 SNPs? Which drug is targeting IFNAR2?

We appreciate this suggestion and have added these plots in Supplementary Figure 16. Many drugs target IFNAR2, most directly, inhaled interferon-β.

Also, have the authors applied any more robust MR method (e.g. median/mode-based estimators)?

In order to minimise horizontal pleiotropy, we have intentionally limited our analysis to locally-acting eQTLs as instrumental variables. The ability to perform more robust Mendelian randomisation methods is contingent on having more than one independent instrument for the exposure. Therefore, we cannot do this analysis without running conditional analysis on both the eQTL study and GenOMICC, or estimating independence using another LD reference (e.g. cojo[5]) which is unlikely to be accurate enough.

Wouldn't protein-> severity be a better MR exercise? There are many studies with available protein QTL summary stats.

Yes, we agree that protein QTL would be valuable instruments for hypothesis generating MR studies (indeed, this has been a focus of some of our recent work[6]). In the present work we have focused on eQTL data because of the abundance of well-powered locally acting eQTL from relevant tissues for use as instruments, but we agree that at careful pQTL MR analysis will be of considerable interest in future work.

Also, using such a small sample as GTEx is prohibitive for statistical power. Have the authors considered using the summary stats of the eQTL-Gen consortium (whole blood, n=32K)?

Thank you for this important improvement. As noted above, we have now used this data to fully replicate our findings (*IFNAR2*: replication p-value 7.46E-04 *TYK2*: replication p-value 5.53E-05). See sec. 2.2.3 for more details.

> Also, why only gene expression was explored as exposure: how about other potential modifiable risk factors (diabetes, FEV, obesity, etc.)?

We agree that this would be an interesting analysis, but for this first report of our findings we prefer to take a conservative approach. As we discuss in sec. 3.2.1, the effect of these risk factors on the development of the critical illness phenotype is relatively small. By using only locally-acting eQTL as instruments, we minimise the risk of falsely inferring causality due to horizontal pleiotropy. This is not possible when using exposure phenotypes further removed from genotype, such as diabetes and obesity. Finally, whilst knowledge of the causal relationships between comorbidities is important, it is less likely to change the overall mortality from Covid-19 than the successful prioritisation of targets for drug repurposing. For example, diabetes is already well-managed in hospitals and intensive care units,[7] and obesity is not modifiable within the timescales necessary to improve survival.

> It is somewhat surprising how MR showed nothing GW significant, but "combined meta-TWAS analysis" revealed 18 genes. Mathematically TWAS and MR (standard IVW without any heterogeneity filtering) are equivalent.

We agree that these are closely-related approaches and in some circumstances can reduce to identical calculations. Fundamentally the difference between them in our application is that we have used single-variant, two-sample MR with cis-eQTL to test a small number of highly-tractable hypotheses with clearly-defined assumptions; in contrast we have used multiple variants for each gene in TWAS as a hypothesis-generating tool. Because our TWAS used multiple variants, with no test for heterogeneity, there is a risk of horizontal pleiotropy - that is, the effect on outcome may not be mediated by the genes identified. This is exemplified by the multiple significant associations in opposite directions in a single locus on chr3: CCR2, CXCL6 and CCR3. Although it is possible, it is unlikely that the effect on outcome is mediated by all three of these genes.

We have amended the discussion to make this caveat clearer to non-expert readers: "it is likely that one, but not all, of these genes is an important mediator of critical illness".

> Is it the combination of blood and lung tissue that made the difference? Figure 3 confirms my confusion: how come the lung-only TWAS shows much weaker signals than the meta-TWAS. It would mean that these are mostly driven by blood-TWAS? Can I see the QQ-plot of the meta-TWAS P-values?

In our meta-TWAS we drew eQTL information from all other tissues to support associations that are nominally significant in lung or blood. We took SNPs that have low p-values ($p<0.01$) in lung or whole blood and looked for common eQTL in all GTEXv8 tissues, in order to increase statistical power, by increasing sample size. Hence, we expected the number of statistically-significant findings to increase. However in response to the reviewer's comment we looked again at a more stringent approach to this analysis.

We have now used fizi (https://github.com/bogdanlab/fizi) to impute summary statistics for ~500K SNPs in order to maximise the overlap between GWAS variants and the eQTL models used for TWAS, and then repeated TWAS for lung and blood and the meta-analysis. This detects only 5 significant associations (Figure 4), in line with the reviewer's expectations, and so we have replaced the TWAS

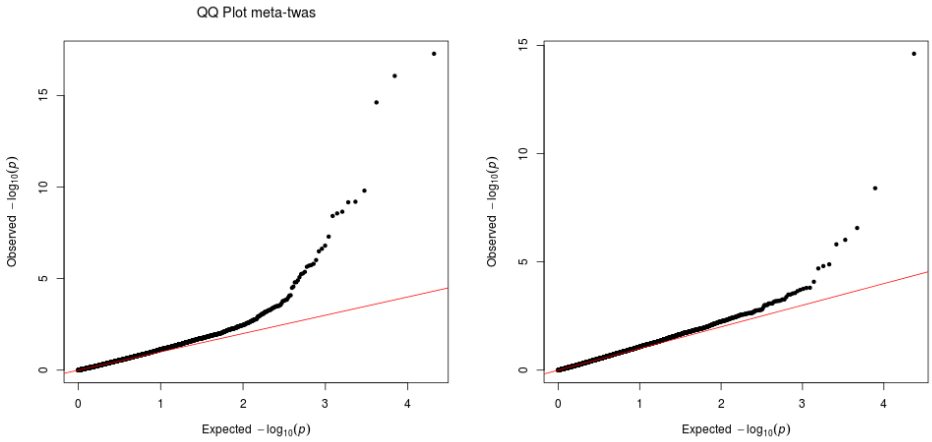


Figure 4: QQ-plot for meta-TWAS (a) original TWAS (b) imputed TWAS.

in the manuscript with this new analysis and revised the methods section. We are grateful to the reviewer for bringing this to our attention.

> The MAIC analysis remains extremely vague: a single (hard-to-read) plot is provided in the supplement, but no exact quantification of statistical evidence of overrepresentation is given, neither significance (or confidence interval) attached to it.

We have expanded this section of the Supplementary Information, including additional explanation of the MAIC methodology, calculations and a link to the systematic review evaluating the evidence from previous studies, and the orginal paper describing the methodology. We used a permutation test to estimate a p-value for enrichment using 1000 iterations with a modelled normal distrubution ($p = 4.2 \times 10^{12}$).

We have replaced the plot with an less-cluttered version, highlighting only the GenOMICC results and connections.

2.2.5 Population structure

> The fact that education and intelligence show the strongest genetic correlation worries me slightly, as it is typically an indicator of uncorrected population stratification and mostly proxies of socio-economic status. Also, how come overall health rating has positive genetic correlation with severe complications?

We agree that the genetic correlations with education, intelligence and obesity may be a reflection of the primary limitation in our study design - the use of opportunistic population controls. We discussed this limitation in the main manuscript but in view of the Reviewer's concern (and similar comments from Reviewers 1 and 3), we have expanded the discussion of this limitation to make this clearer to non-expert readers.

We have performed new validation analyses using two additional test GWAS: GenOMICC vs. 100,000 genomes controls, and the new GenOMICC vs. closely-matched UK Biobank controls. These analyses

did not reveal a significant signal for heritability. We describe these in the main manuscript as follows:

"We were not able to detect a significant signal for heritability in two additional analyses: firstly, using controls from the 100,000 genomes project (in which matching to the GenOMICC cases is less close, which may limit heritability estimation) and secondly, in a second analysis comparing GenOMICC cases UK Biobank controls, with matching of BMI and age where possible. This second analysis was less powerful because of the lack of close matches for 410 cases: n_{cases} = 1260; $n_{controls}$ = 6300 (Supplementary Figure 15)."

Due to the lack of a heritability signal, we were unable to repeat the genetic correlation analysis in these additional GWAS. However, we used genetic covariance to demonstrate that a similar pattern of covariance exists in both of these new analyses: Supplementary Figure 20.

Importantly, the effect of population stratification is extremely unlikely to explain the specific genetic associations we report in Tables 2-4, because these findings are supported by comparison with another external control group (100,000 genomes), and robustly replicate in external data from the HGI meta-analysis and 23andMe.

We now include the following paragraph in the Discussion:

"Because of the urgency of completing and reporting this work, we have drawn controls from population genetic studies with systematic differences in population structure, demographics and comorbid illness, who were genotyped using different technology from the cases. Residual confounding is reflected in the genomic inflation ($\lambda_{0.5}$) value of 1.099 for the primary analysis. We mitigated the consequent risk of false-positive associations driven by genotyping errors by genotyping the majority of our subjects using two different methods (microarray and whole-genome sequencing), and by verifying significant associations using two additional control groups (100,000 genomes and Generation Scotland). The success of these mitigations is demonstrated by robust replication of our sentinel SNPs in external studies. Our meta-analysis, combining GenOMICC with multiple additional sources of genome-wide associations, has a reassuring $\lambda_{0.5} = 1.017$ and confirms the key discoveries from GenOMICC."

> The tissue enrichments look all over the place, I do not see any clear pattern, and barely anything emerges to be significant. Moreover, I do not see the link with spleen/pancreas and COVID19.

We agree that this is not an easily-interpretable result and it is likely to be limited by the statistical power of our GWAS. We report it because some readers will want to see this information, and for a new disease such as Covid-19, these enrichments can contribute to our overall understanding of the pathophysiology. Interestingly, spleen is the only organ (besides the lung) in which striking alterations were discovered in post-mortem studies of patients who died of Covid-19.[8] Also, the expression datasets used are derived from mostly healthy tissues and cannot be expected to reflect gene expression in inflamed tissues during critical illness. Hence, tissue enrichment in spleen may be driven by lymphocyte- or macrophage-expressed genes whose role in disease is executed in other tissues.

> I have the impression that only the EUR subset was used in the entire manuscript, have I missed something? If not, why were the other cases ignored?

We have now clarified this in the results section and in figure legends throughout the manuscript. The non-EUR cases were included where possible but there are limitations on the intpretation of the results of these GWAS. There is considerable inflation in two of the ethnic groups (Supplementary Figure 5), which we is likely to be due to the difficulty in finding well-matched controls in UK biobank.

We have included results from all cases where they contribute meaningfully to the findings, in particular

in the interpretation of LD structure around the replicated sentinel variants (see sec. 1.2.2 for further justification).

I do not see any pathway enrichment analysis in the paper. It guess it has been attempted, but nothing showed up?

We did perform pathway analysis and reported the negative results in Supplementary Table 9. To make this clearer we have now highlighted the description of the pathway analysis methodology in the Materials and Methods section ("Gene-level and pathway analyses").

Given the 1667 sequenced case, could the author compare the frequency of rare variants to those in gnomAD?

We agree that this is an exciting opportunity - we are undertaking a systematic study of rare variants in comparison with both gnomAD and the 100,000 genomes project. This work is now underway but performing it to a high standard is time-consuming, due to the risks of false-positive associations. In view of the potential implications of our GWAS results we have decided to report them immediately, and will report the rare variant analyses seperately when they are completed and replicated.

2.3 Minor comments

1. It has to be explicitly stated what is meant by "equivocal evidence of heterogeneity" for the Heidi test – what heterogeneity P-value they got, have they removed any instruments as a consequence?

In response to this and comments from Reviewer 1, we have now expanded this part of the results section to make this result, and our interpretation of it, clear. Please see our response above: sec. 2.2.4.

2. Regarding SNP-based heritability: please specific on which scale was it estimated. It is never mentioned in the HDL paper either.

We used logistic regression for GWAS analysis of the binary trait: critical illness due to Covid-19. The estimated genetic effects, i.e. log odds ratios, are on the liability scale, which is the scale used for heritability and genetic correlation estimation in HDL. Namely, the scale corresponds to an underlying liability phenotype for severe COVID-19, which follows a logistic distribution and has a variance approximately $\pi^2/3$, or standard deviation (SD) approximately $\pi/\sqrt{3}$.[9]

3. The authors repeatedly use "discovered and replicated", but it is never clearly mentioned how many loci do they claim as new discovery and which ones are replicating previous ones.

We have now clarified this in the replication table, Table 3.

4. "summary statistics 221 from GenOMICC are openly available": A Link would be great there. As it is not on the GenOMICC website.

Many thanks - the link is now live at https://genomicc.org/data and we have added this to the discussion as suggested.

5. Why replication P-values and betas are not listed for 23andMe in Table 3?

We have now included p-values from a meta-analysis of HGI and 23andMe in this table and provide full results in a new supplmentary file.

6. Figure 3 would be more informative to show gcc on top and hgi+23andMe on the bottom in the Miami plot.

We have created a revised version of this plot to show the meta-analysis without GenOMICC, in Supplementary Figure 1.

3 Referee #3 (Remarks to the Author):

3.1 Summary

This manuscript by Baillie and colleagues identifies novel loci and confirms loci associated with critical illness from COVID-19. Using population based controls the authors evaluate the host genetic and transcriptomic associations with severe COVID-19 outcomes requiring intensive care follow up from across more than 200 clinical hospital intensive care sites in the UK. The authors should be credited for acknowledging the heterogeneity in COVID-19 outcomes, and for focusing on one extreme–hospitalized cases. By using an established network of collaboration (and expertise) for critical illness across the UK, they are able to identify a uniform definition of cases.

3.2 Comments

- However, its not clear from the data/table provided that this population of cases is homogeneous. Additional details on the need for mechanical ventilation, or oxygen support, or days (with range) of hospitalization would be helpful to understand if all of these patients are similar, or still represent a spectrum of hospitalized outcomes.

We have now provided additional clinical characterisation of the cohort (Supplementary Figures 8-13). GenOMICC is designed as a pragmatic study to enable research teams to recruit patients quickly in intensive care units during a crisis. For this reason, we obtained minimal clinical data to enable matching of controls, and we recorded additional identifiers to facilitate linkage to other datasets. The additional information in Supplementary Figures 8-13 comes from 1069 GenOMICC cases for whom we have been able to obtain clinical data from the ICNARC Case Mix Programme. This shows that the GenOMICC cohort is largely representative of critically-ill Covid-19 patients across the country. Although this does not prove homogeneity, the features of the GenOMICC cohort are consistent with existing knowledge about critical illness in Covid-19.[10]

We agree that there is heterogeneity in the clinical presentations in Covid-19 - in fact, we have recently reported this using both symptomatology[11] and cytokine measurements (Thwaites et al, manuscript currently undergoing screening at MedRxiv) in the ISARIC 4C project. Importantly, there is some convergence in clinical presentation among the critically-ill group, in whom respiratory failure is the defining feature.[10]

- Similarly, the time frame of recruitment into the study in light of the epidemic in the UK would be helpful. Because, we know globally that those affected prior to governments shutting businesses etc or the adoption of social distancing are different than those later (i.e age), and treatment also evolved with more provider experience.

We have created a new figure (Supplementary Figure 12) which shows that recruitment to the study closely tracked ICU admissions nationwide throughout the first wave of the outbreak in the UK, with a slight delay consistent with our expectations, since patients are not usually recruited to GenOMICC immediately after admission to ICU.

3.2.1 Comorbitity

The one area that the authors address but do not provide details on, are the co-morbidities. We know these are huge risk factors for severe outcomes along with age and sex. * Although the authors indicate in Table 1 the "significant comorbidities" its not clear what that means? In studies out of China and the United States–hypertension is a risk factor. Is that considered in the significant comorbidities? For example, were the majority of comorbidities COPD or Asthma–in which case some of these loci really represent a novel loci for severe hospitalized COPD exacerbation following a viral infection? More details here will also educate us on who was sickest and if genetics goes beyond these comorbidities.

We have provided additional details from data matched with the ICNARC Case Mix Programme in Supplementary Figure 13 and Supplementary Table 2. We did not include hypertension as a risk factor, because we have not recorded this diagnosis in GenOMICC, ISARIC 4C or ICNARC CMP. However, the OpenSafely study provides relevant evidence for the UK population, and found a relatively small hazard ratio for hypertension of 1.09 (95%CI 1.05–1.14), which was mostly mediated by the association with diabetes and obesity.[12]

We previously showed that comorbid illness is a risk factor for death from Covid-19,[13] and for admission to critical care, but the effect sizes are relatively small, effectively excluding the possibility that one of these comorbidities has directly driven the genetic signals we report.[14]

Table 1: Hazard ratios for death for selected comorbidities in ISARIC 4C study.[14]

Comorbidity	Hazard ratio (95% confidence interval)
Obesity	1.33 (1.19-1.49)
Diabetes	1.06 (0.99-1.14)
Chronic cardiac disease	1.16 (1.08-1.39)
Chronic pulmonary disease	1.17 (1.09-1.27)

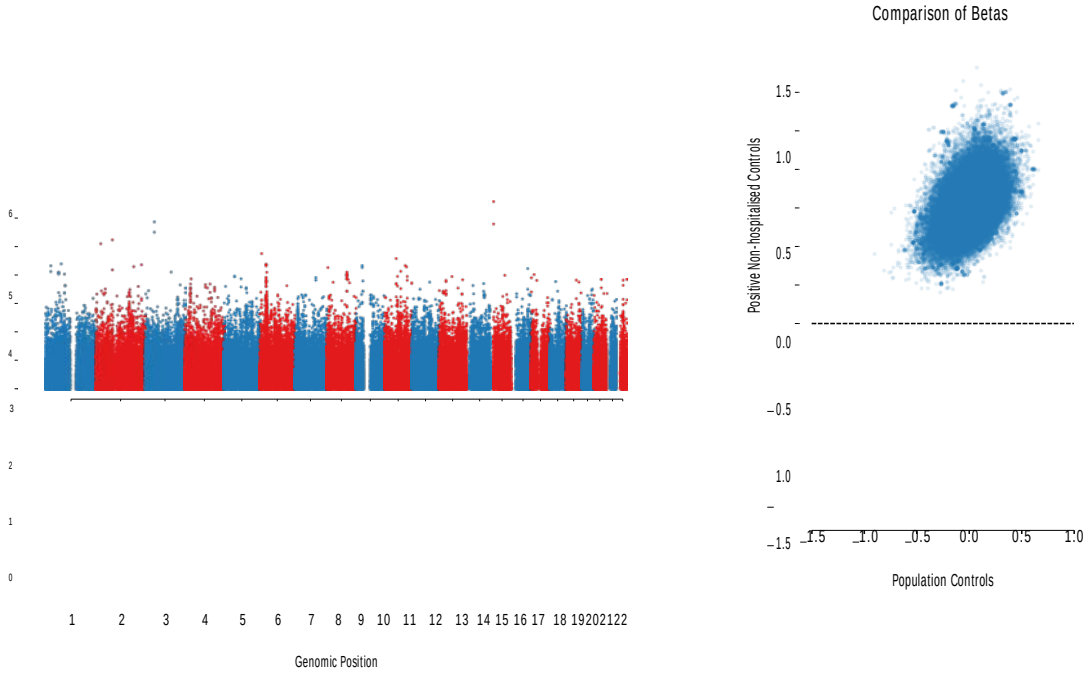
Baillie and colleagues identify the data rich UK Biobank as a source of population based controls. They select 5 controls per case. * However, details on these "controls" with regards to age, comorbidities and sex should be provided in a table. If they were younger and healthier it does suggest that they would have been unlikely anyway to be in the intensive care units at least in the early days of the pandemic. And what about obesity? Was this matched on?

We did not match on obesity in our initial analysis, but have now completed a repeat GWAS in which we sampled UK Biobank controls from a BMI- and age- distribution matching the GenOMICC cases

(this is described in full in sec. 3.2.2).

- Interestingly, the authors have access to COVID-19 positive individuals in the UK Biobank, but chose to exclude them. What was the rationale? This group of individ- uals would be the best match, and ideally should be considered since they were exposed, and did not require hospitalization. If we are trying to identify a severity locus–this would be the best approach.

We agree that, with sufficient numbers, this will be the best design. However, at present the numbers are limited. In response to this query we have completed a repeat GWAS comparing GenOMICC cases to UK Biobank Covid-positive cases (1676 case in EUR group) who were not hospitalised (562 controls), as suggested. Although the additional results are concordant with the primary analysis reported, the power is insufficient to detect genome-wide significant associations.



(a) Results of comparison with 562 Covid-positive non-hospitalised controls from UK Biobank (b) Comparison with primary analysis.

- Using population based controls in the manner of 5:1 may increase sample size but the authors should acknowledge the underlying assumption is that they believe that these population controls may be exposed but not require hospitalization. Otherwise there is inherent misclassification of the controls.

We agree that there is some inevitable misclassification because the entire UK population has not been exposed to SARS-CoV-2. This contaminates the control group with individuals who may exhibit the critical illness phenotype if exposed, but in whom it has not yet been unmasked by exposure to the virus. Of course, the effect of this will be to bias any signals towards the null. We have added discussion of this point in the results section:

"Ancestry-matched controls were selected from the large population-based cohort UK Biobank in a ratio of 5 controls to 1 case. Controls with a known positive Covid-19 test were excluded. The inevitable presence of individuals in the control group, who may exhibit the critical illness phenotype if exposed to SARS-CoV-2, is expected to bias any associations towards the null."

- The authors also only present in the main figures the European plots and locus zooms (not all locus zooms look convincing). Given the worldwide pandemic that affects all ancestry groups, its seems crucial that the genetics for all groups is also presented in the main portion of this paper. It is a strength of this study. And it could also aid our understanding of the Chromosome 3 locus– if credible sets were used to narrow down this region. This could be done (or attempted) with this diverse population and aid the overall genetic community.

We agree this is important and we have now replaced the locus zoom plots in Figure 1 and in Supplementary Figure 7 with plots showing LD patterns for each ancestry group.

- The PCA plots supplementary should also show each ancestry separately with controls so that we can see the overall distributions, and more than just PC1 and PC3.

We have now included a more comprehensive series of PCA plots reporting PCs for all ethnicities in Supplementary Figure 3. Reviewer 1 asked for the same thing so we have discussed this further above (sec. 1.2.1) and reproduced the plots in Figure 1

The addition of the transcriptomics and gene grouping is interesting and seems adequate in approach. The heritability analysis is likely lacking sufficient power and that should be noted as a limitation and not reflective of a lack of heritability given the effect sizes and that this is SNP based. But the heritability correlation is very intriguing.

We have changed the text to reflect this, as follows: "We used the high-definition likelihood (HDL) method[15] to provide an initial estimate the SNP-based heritability, that is the proportion of phenotypic variance that is captured by additive effects at common SNPs, to be 0.065 (SE = 0.019) for severe Covid-19. Including rare variants in future analyses, with larger numbers of cases, will provide a more comprehensive estimate of heritability."

3.2.2 Genetic correlations

- The authors cite correlations with education and intelligence and with body mass measurements. These alone are reasons why the cases/controls should be matched on these factors, or ideally a propensity score matching approach should be done. We know that different people were exposed at different times based on their occupation, public trans- portation, etc and we need to address those differences. Similarly, we know BMI is a risk factor for severe outcomes, so this should be addressed in the analysis. The authors do excellent work to identify these measures but now must be sure that they are not influencing or confounding their genetic association also.

We agree - as we discuss in the manuscript, and has been pointed out by other reviewers, the effect of selection bias in UK Biobank is a significant caveat in the interpretation of the genetic correlations. Firstly, we have expanded the discussion of limitations of the study in order to make this clear (see sec. 2.2.5). Secondly, in response to the queries from all reviewers about matching of controls in UK Biobank, population structure and residual inflation, we have repeated the GWAS using GenOMICC cases and matched controls.

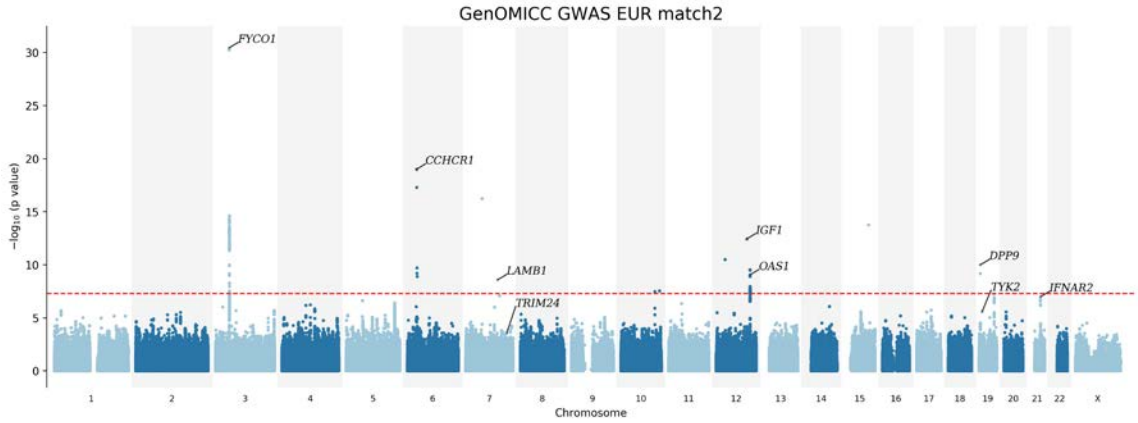
We include the following description in the Supplemental Information:

”Because of the evidence of residual inflation in the GenOMICC EUR UK Biobank analysis, and the genetic correlations with obesity and educational attainment (Supplementary Figure 20), we undertook further analysis to examine the effect of additional correction for population structure. We performed a GWAS in which we restricted the analysis to cases for whom UK Biobank controls could be identified according to the following rules:

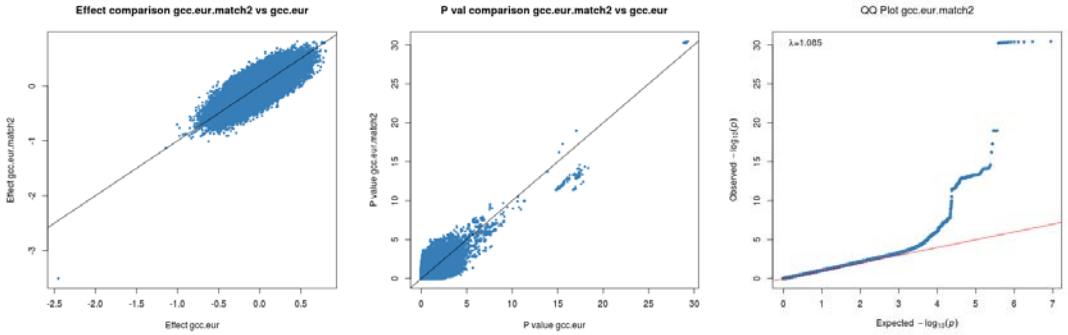
- individual matches by ancestry, sex, age, and deprivation quintile
- BMI sampled from a distribution that parallels the ICNARC CMP BMI distribution for the GenOMICC cases (Supplementary Figure 10)

Applying these rules produced a smaller comparison than the primary analysis: n_{cases} = 1260; $n_{controls}$ = 6300.”

Results can be seen in Supplementary Figure 15 and are reproduced here for convenience: Figure 5.



(a)



(b) (c) (d)

Figure 5: Results from secondary analysis with matched controls. (a) Manhattan plot showing single nucleotide polymorphism (SNP)-level p-values for genome-wide significant associations (red horizontal line shows genome-wide significance at $-log_{10}(5 \times 10^{-8})$). (b) Correlation with primary analysis in effect sizes (β), and p-values (c). (d) QQ plot.

Importantly, despite the stringent matching, inflation is still present in this new analysis (Matched control analysis: $\lambda 0.5 = 1.085$; Primary analysis $\lambda 0.5 = 1.099$). This comes at a cost of more than 416 cases removed from the analysis. Nonetheless, the core discoveries we report in Table 2 are still evident in this new analysis, above or close to the genome-wide significance threshold (Figure 5a).

Overall, this is an easy paper to read and is clear in writing style. Many different genomics approaches were employed, but the overall message is that new loci have been identified. To solidify those loci, additional epidemiologic and clinical descriptions and inclusion in the analysis are necessary.

Minor note: some of the references are not correctly annotated (or there is an issue with how they have printed–with extra characters in the citation).

We have corrected this in the revised manuscript.

References

1. Horby, P.W., Mafham, M., Bell, J.L., Linsell, L., Staplin, N., Emberson, J., Palfreeman, A., Raw, J., Elmahi, E., Prudon, B., Green, C., Carley, S., Chadwick, D., Davies, M., Wise, M.P., Baillie, J.K., Chappell, L.C., Faust, S.N., Jaki, T., Jefferey, K., Lim, W.S., Montgomery, A., Rowan, K., Juszczak, E., Haynes, R. & Landray, M.J. LopinavirRitonavir in patients admitted to hospital with COVID-19 (RECOVERY): A randomised, controlled, open-label, platform trial. *The Lancet* **0**, (2020).

2. Horby, P., Lim, W.S., Emberson, J.R., Mafham, M., Bell, J.L., Linsell, L., Staplin, N., Brightling, C., Ustianowski, A., Elmahi, E., Prudon, B., Green, C., Felton, T., Chadwick, D., Rege, K., Fegan, C., Chappell, L.C., Faust, S.N., Jaki, T., Jeffery, K., Montgomery, A., Rowan, K., Juszczak, E., Baillie, J.K., Haynes, R. & Landray, M.J. Dexamethasone in hospitalized patients with covid-19 - preliminary report. *The New England journal of medicine* (2020).doi:10.1056/NEJMoa2021436

3. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M. & Yang, J. Integration of summary data from gwas and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**, 481–7(2016).

4. Battle, A., Brown, C.D., Engelhardt, B.E. & Montgomery, S.B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213(2017).

5. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., Frayling, T.M., McCarthy, M.I., Hirschhorn, J.N., Goddard, M.E. & Visscher, P.M. Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. *Nature genetics* **44**, 369–75, S1–3(2012).

6. Bretherick, A.D., Canela-Xandri, O., Joshi, P.K., Clark, D.W., Rawlik, K., Boutin, T.S., Zeng, Y., Amador, C., Navarro, P., Rudan, I., Wright, A.F., Campbell, H., Vitart, V., Hayward, C., Wilson, J.F., Tenesa, A., Ponting, C.P., Baillie, J.K. & Haley, C. Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLOS Genetics* **16**, e1008785(2020).

7. Rhodes, A., Evans, L.E., Alhazzani, W., Levy, M.M., Antonelli, M., Ferrer, R., Kumar, A., Sevransky, J.E., Sprung, C.L., Nunnally, M.E., Rochwerg, B., Rubenfeld, G.D., Angus, D.C., Annane, D., Beale,
R.J., Bellinghan, G.J., Bernard, G.R., Chiche, J.-D., Coopersmith, C., De Backer, D.P., French, C.J., Fujishima, S., Gerlach, H., Hidalgo, J.L., Hollenberg, S.M., Jones, A.E., Karnad, D.R., Kleinpell, R.M., Koh, Y., Lisboa, T.C., Machado, F.R., Marini, J.J., Marshall, J.C., Mazuski, J.E., McIntyre,
L.A., McLean, A.S., Mehta, S., Moreno, R.P., Myburgh, J., Navalesi, P., Nishida, O., Osborn, T.M., Perner, A., Plunkett, C.M., Ranieri, M., Schorr, C.A., Seckel, M.A., Seymour, C.W., Shieh, L., Shukri, K.A., Simpson, S.Q., Singer, M., Thompson, B.T., Townsend, S.R., Van der Poll, T., Vincent, J.- L., Wiersinga, W.J., Zimmerman, J.L. & Dellinger, R.P. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. *Intensive Care Medicine* **43**, 304–377(2017).

8. Dorward, D.A., Russell, C.D., Um, I.H., Elshani, M., Armstrong, S.D., Penrice-Randal, R., Millar, T., Lerpiniere, C.E., Tagliavini, G., Hartley, C.S., Randall, N.P., Gachanja, N.N., Potey, P.M., Anderson, A.M., Campbell, V.L., Duguid, A.J., Qsous, W.A., BouHaidar, R., Baillie, J.K., Dhaliwal, K., Wallace, W.A., Bellamy, C.O., Prost, S., Smith, C., Hiscox, J.A., Harrison, D.J., Lucas, C.D. & ICECAP Tissue-specific tolerance in fatal Covid-19. *medRxiv*

2020.07.02.20145003(2020).doi:10.1101/2020.07.02.20145003

9. Pawitan, Y., Seng, K.C. & Magnusson, P.K.E. How many genetic variants remain to be discovered?
PloS one **4**, e7969(2009).

10. Docherty, A.B., Harrison, E.M., Green, C.A., Hardwick, H.E., Pius, R., Norman, L., Holden, K.A., Read, J.M., Dondelinger, F., Carson, G., Merson, L., Lee, J., Plotkin, D., Sigfrid, L., Halpin, S., Jackson, C., Gamble, C., Horby, P.W., Nguyen-Van-Tam, J.S., Ho, A., Russell, C.D., Dunning, J., Openshaw, P.J., Baillie, J.K. & Semple, M.G. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study. *BMJ* **369**, (2020).

11. Millar, J.E., Neyton, L., Seth, S., Dunning, J., Merson, L., Murthy, S., Russell, C.D., Keating, S., Swets, M., Sudre, C.H., Spector, T.D., Ourselin, S., Steves, C.J., Wolf, J., Investigators, I., Docherty, A.B., Harrison, E.M., Openshaw, P.J., Semple, M.G. & Baillie, J.K. Robust, reproducible clinical patterns in hospitalised patients with COVID-19. *medRxiv* 2020.08.14.20168088(2020).doi:10.1101/2020.08.14.20168088

12. Williamson, E.J., Walker, A.J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C.E., Curtis, H.J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H.I., MacKenna, B., Tomlinson, L., Douglas, I.J., Rentsch, C.T., Mathur, R., Wong, A.Y.S., Grieve, R., Harrison, D., Forbes, H., Schultze, A., Croker, R., Parry, J., Hester, F., Harper, S., Perera, R., Evans, S.J.W., Smeeth, L. & Goldacre, B. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436(2020).

13. Knight, S.R., Ho, A., Pius, R., Buchan, I., Carson, G., Drake, T.M., Dunning, J., Fairfield, C.J., Gamble, C., Green, C.A., Gupta, R., Halpin, S., Hardwick, H.E., Holden, K.A., Horby, P.W., Jackson, C., Mclean, K.A., Merson, L., Nguyen-Van-Tam, J.S., Norman, L., Noursadeghi, M., Olliaro, P.L., Pritchard, M.G., Russell, C.D., Shaw, C.A., Sheikh, A., Solomon, T., Sudlow, C., Swann, O.V., Turtle, L.C., Openshaw, P.J., Baillie, J.K., Semple, M.G., Docherty, A.B. & Harrison, E.M. Risk stratification of patients admitted to hospital with covid-19 using the isaric who clinical characterisation protocol: Development and validation of the 4C mortality score. *BMJ (Clinical research ed.)* **370**, m3339(2020).

14. Docherty, A.B., Harrison, E.M., Green, C.A., Hardwick, H.E., Pius, R., Norman, L., Holden, K.A., Read, J.M., Dondelinger, F., Carson, G., Merson, L., Lee, J., Plotkin, D., Sigfrid, L., Halpin, S., Jackson, C., Gamble, C., Horby, P.W., Nguyen-Van-Tam, J.S., Ho, A., Russell, C.D., Dunning, J., Openshaw, P.J., Baillie, J.K. & Semple, M.G. Features of 200.167em133 uk patients in hospital with covid-19 using the isaric who clinical characterisation protocol: Prospective observational cohort study. *BMJ (Clinical research ed.)* **369**, m1985(2020).

15. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nature Genetics* **52**, 859–864(2020).

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

I thank the authors for careful consideration of my comments. I am satisfied with the responses and have no additional comments.

Referee #2 (Remarks to the Author):

The authors have very carefully addressed my comments. While I'm generally satisfied with the new version, two small points may need clarifications.

Re: "2.2.2 Replication": While I am happy with the authors effort to ensure that no artefacts have crept in due to the suboptimal design, still haven't received any answer for the reason of non-

replication of 3/8 hits. I think the readers would be curious to hear the authors guesses.

2.2.4: The eQTL-Gen data set is large enough to lend itself to approximate conditional analysis. I do not see justified the authors statement about the inaccuracy of COJO, there is large enough reference panel of EUR samples. MR based on 2-3 variants is extremely vulnerable to pleiotropy – I'd be more convinced to see how MR works with more instruments coming from a conditional analysis based on the eQTL-Gen summary stats.

Referee #3 (Remarks to the Author):

The authors have addressed all of the comments and made extra efforts to add analyses to address points raised.
The two remaining concerns are:

1) The inflation factors. These remain inflated and when looking at the PCA plots it looks like there may be some substructure or a European distinction in this population that could be contributing to this inflated lambda (visible on PC4/5, PC 3/4).
When you restrict to just Europeans and their matched controls it appears that the matching may not be so great for the Europeans but its hard to distinguish because of the overlap of circles. Did you consider increasing the number of PCA's? I would consider running with more to see if it improves, and at least reporting that increasing the PC's did not alter the inflation. Because then it may be the genotyping platform difference as you describe.

The East Asian and South Asian plots are just too inflated to have reliable results and this needs to be addressed in the text.

2) The forest plots detailing the different ancestral populations. At least for two regions: (rs2109069) and rs10735079) the odds ratios are not consistent across populations and go in different directions. For the DPP9 locus this might suggest a spurious association given that there are only two SNPs indicated in the zoom plot as associated with limited recombination in the region. At the least I would address in the text that these regions were not fully replicated across populations and this could be because of true lack of replication or because of confounding, or sample size. Although they were replicated in HGI and 23&me. Please also detail if the 23&me and hgI populations included are all European with European controls or multi-ethnic or even multi-ancestry across Europe.

Author Rebuttals to First Revision:

1 Referee #1

I thank the authors for careful consideration of my comments. I am satisfied with the responses and have no additional comments.

2 Referee #2

The authors have very carefully addressed my comments. While I'm generally satisfied with the new version, two small points may need clarifications.

Re: "2.2.2 Replication": While I am happy with the authors effort to ensure that no artefacts have crept in due to the suboptimal design, still haven't received any answer for the reason of non-replication of 3/8 hits. I think the readers would be curious to hear the authors guesses.

Since all three variants are in the MHC region on chromosome 6, we believe this may be due to residual population structure, the play of chance, or a failure of the replication studies to detect these signals. In response we have added the following sentence to clarify the presentation of the results: "Three variants, all in a region of the genome in which population stratification is difficult to control (the major histocompatibility complex, MHC), did not replicate. Further studies will be required to determine whether these associations are real."

2.2.4: The eQTL-Gen data set is large enough to lend itself to approximate conditional analysis. I do not see justified the authors statement about the inaccuracy of COJO, there is large enough reference panel of EUR samples. MR based on 2-3 variants is extremely vulnerable to pleiotropy

– I'd be more convinced to see how MR works with more instruments coming from a conditional analysis based on the eQTL-Gen summary stats.

We are grateful for the reviewer pushing us to improve this analysis. We have completed the analysis as suggested and generated compelling validation. We have described this analysis in the following section of text in Materials and Methods:

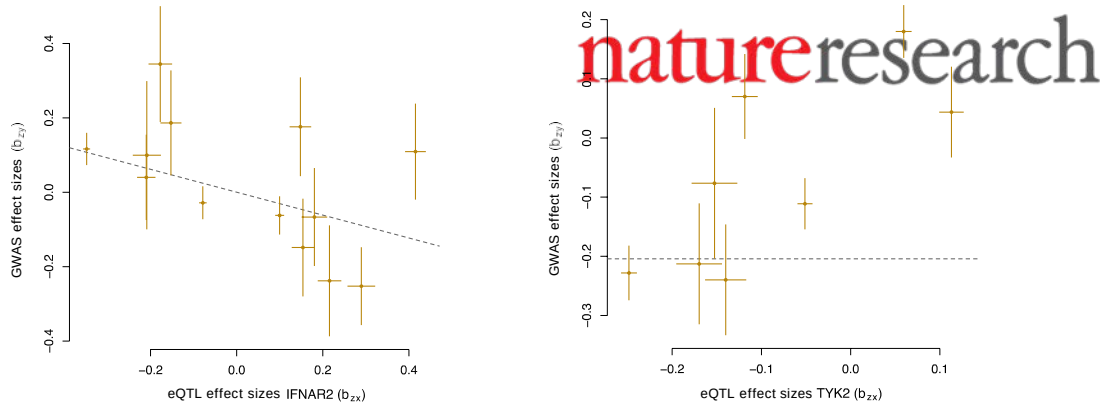
In order to further validate the analyses above, Generalized Summary-data-based Mendelian Randomization (GSMR)[1] was performed using exposure data from https://www.eqtlgen.org (accessed 01/10/20)[2] and the GenOMICC EUR data for TYK2 and IFNAR2. GSMR was performed using GCTA version 1.92.1 beta6 Linux. Pleiotropic SNPs were filtered using HEIDI-outlier test (threshold $= 0.01$) and instrument SNPs were selected at a genome-wide significance level ($PeQTL <$ 5e-8) using LD clumping (LD^{r2} threshold $= 0.05$). The imputed genotypes for 50,000 unrelated individuals (based on SNP-derived genomic relatedness < 0.05 using HapMap 3 SNPs) from UK Biobank were used as the LD reference for clumping. GSMR accounts for remaining LD not removed by the clumping analysis.

The results are strongly significant and are reported in the Supplementary Information. We have reproduced them here in Supplementary Table 7 and Figure 1 for convenience.

In order to further validate the key Mendelian randomisation findings, generalized summary-data-based Mendelian Randomization (GSMR)[1] was performed using multiple independent (Methods) SNPs for IFNAR2 and TYK2. Using data from eQTLgen,[2] 13 and 8 independent SNPs after HEIDI-outlier test were identified for each gene respectively. The GSMR results replicated the SMR results, with both a consistent direction-of-effect and a significant p-value for both genes(Supplementary Table 7, Figure 1).

Table 1: Results of multiple independent instruments Mendelian randomisation using GenOMICC GWAS EUR as outcome, for IFNAR2 and TYK2. - effect size; se - standard error for ; p - Mendelian randomisation p-value; nsnp_HEIDI - number of SNPs included.{#tbl:mr.cojo}

Exposure	β_{xy}	se	p	nsnp_HEIDI
IFNAR2	-0.307834	0.115121	0.0075	13
TYK2	0.74874	0.168939	9.33e-06	8



(a) (b)

Figure 1: $\beta_{exposure}$ vs $\beta_{outcome}$ plots for each variant considered for Mendelian randomisation. Additional analysis using SNPs from eQTLgen is shown in (c) for IFNAR2 and (d) for TYK2.

3 Referee #3

The authors have addressed all of the comments and made extra efforts to add analyses to address points raised. The two remaining concerns are:

1) The inflation factors. These remain inflated and when looking at the PCA plots it looks like there may be some substructure or a European distinction in this population that could be contributing to this inflated lambda (visible on PC4/5, PC 3/4). When you restrict to just Europeans and their matched controls it appears that the matching may not be so great for the Europeans but its hard to distinguish because of the overlap of circles. Did you consider increasing the number of PCA's? I would consider running with more to see if it improves, and at least reporting that increasing the PC's did not alter the inflation. Because then it may be the genotyping platform difference as you describe.

As suggested, we repeated the primary analysis using 20 PCs as covariates and found no improvement in the inflation. We have added the following comment in the Supplementary Information: "Repeating

the analysis using more principal components (20PCs) as covariates did not improve the inflation ($\lambda 0.5 = 1.10$)."

> The East Asian and South Asian plots are just too inflated to have reliable results and this needs to be addressed in the text.

We have added the following sentence in the main paper: "There was a high level of residual inflation in the South Asian and East Asian ancestry groups, rendering results in these subgroups unreliable (Supplementary Figure 5)."

> 2) The forest plots detailing the different ancestral populations. At least for two regions: (rs2109069 and rs10735079) the odds ratios are not consistent across populations and go in different directions. For the DPP9 locus this might suggest a spurious association given that there are only two SNPs indicated in the zoom plot as associated with limited recombination in the region. At the least I would address in the text that these regions were not fully replicated across populations and this could be because of true lack of replication or because of confounding, or sample size. Although they were replicated in HGI and 23&me. Please also detail if the 23&me and hgI populations included are all European with European controls or multi-ethnic or even multi-ancestry across Europe.

We have altered the text as suggested to be clear that the signals were not consistent across all ancestry groups: "Observed heterogeneity in effect size may be due to genuine differences between ancestry groups, or due to the limited statistical power in smaller groups (evident from the broad confidence intervals), or due to residual confounding."

Reassuringly, the 95% confidence intervals for all groups overlap. We looked more deeply at both regions in order to establish whether there is support from nearby variants in LD with the lead SNP. In the region of rs2109069 (DPP9), only one significant SNP in LD with the lead SNP reported in our primary analysis. Several other SNPs were removed during QC and filtering - we found that many of these were removed by the genotype call rate > 0.99 filter, but in many cases the genotype call rate was very high. We have plotted the relationship between *pGWAS* and r^2 with the lead SNP for these variants in Figure 2, which provides additional support from nearby genotyped variants.

Because they are all close to 99% (between 97% and 99%) I removed the filter and ran a gwas for these snps, and I found that the ld pattern is followed quite well. I am attaching a plot of -log10(P) vs R2 and as you can see the p-value decreases with r2 as you can expect

We have added the following clarifications to the main paper:

"Since no study of critical illness in Covid-19 of sufficient size is available, replication was sought in a meta-analysis of data from 2415 hospitalised Covid-19 cases and 477741 population controls from the Covid-19 Host Genetics Initiative (HGI, **mixed ancestry**, with UK Biobank cases and controls excluded) and 1128 cases and 679531 controls in the 23andMe Inc"broad respiratory phenotype" **(EUR ancestry)**, which includes cases reported being placed on a ventilator, being administered oxygen, or having pneumonia versus controls who did not report positive tests."

and to the Materials and Methods section:

"The 23andMe study comprises cases and controls from EUR genetic ancestry group. The HGI B2 analysis is a trans-ancestry meta-analysis, with the great majority of cases being multi-ethnic European (EUR and FIN), with 238 cases of non-European ancestry (176 Admixed American, AMR, from BRACOVID study and 62 South Asian, SAS, from the GNH study)."

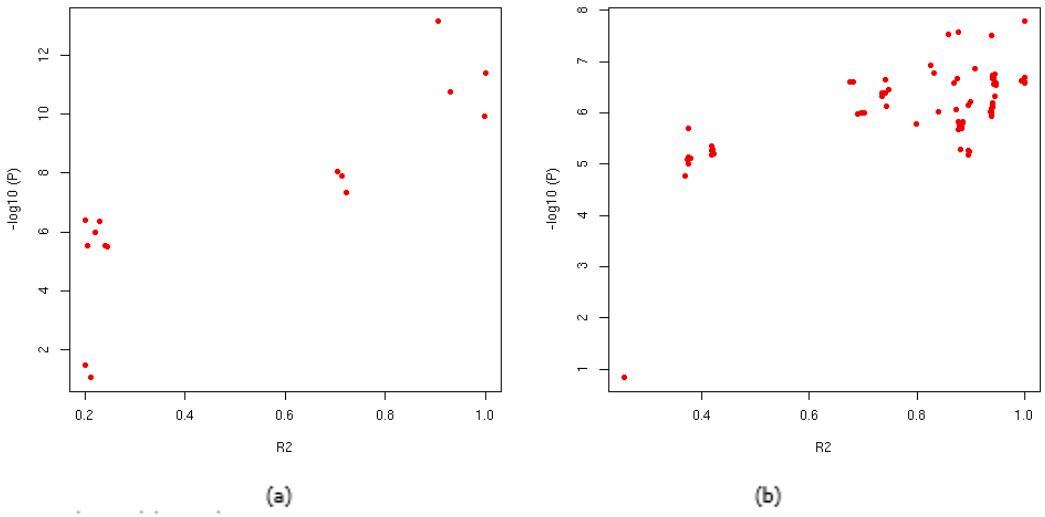


Figure 2: Scatterplots showing $_{GWAS}$ and r^2 (b) for SNPs in LD with (a) rs2109069 and (b) rs10735079.

References

1.Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M.R., McGrath, J.J., Visscher, P.M., Wray, N.R. & Yang, J. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nature communications* **9**, 224(2018).

2.Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg-Guzman, J., Kettunen, J., Powell, J., Lee, B., Zhang, F., Arindrarto, W., Beutner, F., Consortium, B., Brugge, H., Consortium, i., Dmitreva, J., Elansary, M., Fairfax, B.P., Georges, M., Heijmans, B.T., Kähönen, M., Kim, Y., Knight, J.C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U.M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Pritchard, J., Raitakari, O., Rotzschke, O., Slagboom, E.P., Stehouwer, C.D.A., Stumvoll, M., Sullivan, P., Hoen, P.A.C.', Thiery, J., Tönies, A., Dongen, J. van, Iterson, M. van, Veldink, J., Völker, U., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G.W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., Meurs, J. van, Prokisch, H., Ahsan, H., Pierce, B., Lehtimäki, T., Boomsma, D., Psaty, B.M., Gharib, S.A., Awadalla, P., Milani, L., Ouwehand, W., Downes, K., Stegle, O., Battle, A., Yang, J., Visscher, P.M., Scholz, M., Gibson, G., Esko, T. & Franke, L. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367(2018).doi:10.1101/447367

Reviewer Reports on the Second Revision:

Referee #2 (Remarks to the Author):

I thank the authors for having reassuringly addressed all my remaining criticism. I am now convinced of the results of this paper.

Referee #3 (Remarks to the Author):

The authors have addressed all comments. Although some comments don't have answers (i.e. inflation) the caveats included in the text should provide guidance and information to the readers when making inferences.