

Peer Review File

Manuscript Title: The major genetic risk for severe COVID-19 is inherited from Neandertals

Editorial Notes:

Redactions – Third Party Material

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referee #1 (Remarks to the Author):

Report on 'The major genetic risk factor for severe COVID-19 is inherited from Neandertals'

This manuscript investigates the possibility that a region on chromosome 3 that was recently associated with risk for severe Covid-19 outcome has been inherited from Neandertals. The authors compare the specific variants in the associated region with different Neandertal genomes as well as genetic data from several present day human populations. They find that many variants that are in strong linkage to the top Covid-19-associated variant are also found (in homozygosity) in a Neandertal genome from the Vindija cave, a pattern also supported by further Neandertal genomes. The authors conclude that this haplotype was indeed inherited from Neandertals, and support their conclusion with simulations of shared ancestry and by phylogenetic analysis of the haplotypic region across modern human populations.

The analysis is obviously very timely, and a better understanding of the genetic risk variants and their distribution is of very broad interest to the scientific community, but also of particular importance for any public health efforts in the face of this pandemic. The evolutionary perspective taken here is particularly intriguing.

The hypothesis is straight-forward and plausible. Inheritance of Neandertal DNA in modern humans is well established, and this might be a particular interesting case. The overall analysis and presentation of the data and results appears sound and I can identify no major flaws, but there is some information missing that would improve the accessibility and interpretability of the results. I have detailed these points below.

I am a bit concerned about the use of the Covid19hg data. I know that this data is made available for immediate research use, but I am not sure whether the publication of this data is also 'free', before it has been published by the Covid19hg initiative. The section on the covid19hg website about 'Terms of use of results and authorship' is still empty, stating only a note 'Coming soon'. So I would urge the authors to obtain a letter from the coordinators of Covid19hg stating that they approve the publication of these results, or provide other verifiable proof that the publication of these results does not represent misconduct.

I would find it important to see the general frequency of Vindija 33.19 Neandertal ancestry along the chromosome 3 in Europeans (this is where most subjects in this data are from), in order to judge the pattern observed in the Covid-19-associated region. If the frequency of Neandertal ancestry is generally low, this would highlight the particular pattern observed for the Covid-19-associated locus.

At the end of the first paragraph (main text) the authors state that the association with the ABO locus was not replicated by the Covid19hg data. However, this might simply be due to the fact that the primary endpoints of the two studies differ ('severe respiratory failure' vs. 'hospitalization'). So I suggest adding something like the following note to this sentence: ", possibly due to different case definitions between the studies". This will not lower the authors' point that the chr3 locus is the major association with Covid-19, but it will prevent the notion that the earlier study might have yielded erroneous results.

I do not agree that positive selection at this region of interest is unlikely. Clearly, the genes in this region (which include cell-surface receptors that are often targeted by viruses to enter host cells) have more functions/phenotypes than just the effect on Covid-19 risk, so it is certainly possible that variants in this region were selected for other reasons, but now one (or more) of the variants yield a disadvantage in the face of Covid-19. The variants under positive selection and those now associated with Covid-19 could also be different and just be linked. Such effects of genetic draft

are well described for the HLA region (Lenz et al. 2016 MBE, doi:10.1093/molbev/msw127). In fact, this would be a good explanation why this haplotype was maintained in modern humans in the first place, so I would urge the authors to include this as a possibility, especially at the end of the discussion, where they speculate about the role of this haplotype for the health of Neandertals.

Along those lines, it should also be stated more clearly in the text, that the model-based rejection of the possibility for this haplotype to derive from a common ancestor is based on the assumption of neutral evolution (or so I assume). I guess that a model with balancing selection on this haplotype would yield a different result.

The shared red color between Fig1A and 1B suggests similar effects, but Fig1B does not actually show association with Covid-19, it only shows linkage to the lead SNP in Fig.1A. I find this somewhat misleading and thus suggest to use a different color to indicate LD patterns in Fig1B.

The fact that the association signal extends (at a lower level) to 400-500 kb is brushed over with a short sentence in the legend of figure 1, stating that "some individuals carry a Neandertal-like haplotype that extends to ~400 kb". It is not stated how this information was obtained nor how it was verified. I think this needs to be explained in more detail, as it suggests that it is not only one haplotype that drives the association with Covid-19, as could be understood from the description of 'the Neandertal haplotype' throughout the text. I guess this signal comes from recombination/haplotype decay, which is in agreement with the scenario laid out here, but it should be described more clearly.

In the main text, the authors mainly focus their analyses on the SNPs that are in very high LD ($R^2 > 0.99$) with the top SNP from Ellinghaus et al. 2020. However, in Fig1B, to which the main text refers for these SNPs, all SNPs with an $R^2 > 0.1$ are marked in the same color (red), creating the impression that all red SNPs are in very high LD. I find this somewhat misleading and think that it would be helpful if these top SNPs (with $R^2 > 0.99$) were colored differently in Fig. 1B, so that information is conveyed more transparently.

In Fig.2, it is not clear why there are multiple haplotypes in the shaded box. In the text it is stated that these haplotypes are associated with Covid-19, but it is not clear how this was determined. Do they all carry the top associated variant? This is a bit confusing as otherwise in the text mainly one haplotype is mentioned (see also Fig. 3). Please clarify. Also, how robust is this clustering result with regard to the chosen substitution model, e.g. how would it look like with Tamura-Nei model?

In Fig.3, the frequency of the Neandertal haplotype is shown across 1KG populations. However, it is not clear how the data and the populations were selected for this analysis (the 1KG data contains more populations than what is shown here). This analysis needs to be described in more detail. I also find the choice of symbols for the populations where the haplotype is missing suboptimal, as they give much more weight to 'missingness' than to very low frequency.

What is 'the Neandertal risk variant', described in the population frequency analysis? Is it the top-associated variant from Ellinghaus et al. 2020 or the one described in Fig. S1 as being in perfect LD with that top SNP or something entirely different. This should be more clearly described in the text.

Referee #2 (Remarks to the Author):

This study from Zeberg and Paabo tests the hypothesis that major genetic risk factor for severe COVID-19 on chromosome 3 was introduced by Neanderthal introgression into modern humans. Given that relatively little is known about the host genetics associated with COVID-19, this is a potentially exciting result that has major implications for both our understanding the host genetic

reponse to COVID-19 and to the genetic impact of Neanderthal introgression.

The main lines of evidence offered to support a Neanderthal origin for the risk haplotype on chromosome 3 (following techniques used by Huerta-Sanchez and colleagues in their article to show that the variant at EPAS1 that conferred adaptation to high-altitudes was introduced by Denisovan introgression) are that:

- 1) the risk haplotype for COVID-19 is genetically closer to the Neanderthal haplotypes (specifically the Vindija Neanderthal) relative to other haplotypes segregating in modern humans
- 2) the risk haplotype is too long to have been an ancestral polymorphism that was inherited unrecombined across human populations, and
- 3) the geographic distribution of the risk haplotype wherein the risk haplotype is at appreciable frequencies in south Asia, at lower frequencies in Europe but is absent in Africa.

While the first line of evidence is convincing, the second and third lines are less so.

1. The authors argue that the Neanderthal risk haplotype is almost completely absent in Africa. This statement contradicts Ellinghaus et al. 2020 (Figure S8) that shows substantial frequencies of the risk allele in African populations with frequencies ranging from ~2% in Sierra Leone to ~7% in Nigeria and Kenya (comparable to the frequencies seen in Europe and higher than those in East Asia).

Thus, the authors need to clarify this contradiction.

It is plausible, however, that an introgressed haplotype could have attained high enough frequencies in Africa through gene flow into Africa from non-African populations. To show that the risk haplotype is indeed introgressed, the authors should inspect recent maps of introgression (e.g Chen et al. 2020). Such an analysis could further clarify the evolutionary history of the risk haplotype in modern human populations.

2. An analysis of the length of the haplotype is used to argue against the possibility of incomplete lineage sorting producing the patterns we see today. This analysis uses a recombination of 0.53 cM/Mb.

What recombination map was used in this analysis ? Given the uncertainties of recombination maps at kilobase scales, it would be essential to redo this analysis using either admixture-based and/or pedigree-based maps.

Referee #3 (Remarks to the Author):

The authors find the recently identified COVID-19 risk locus in Chromosome 3 has a haplotype structure that can be explained by introgression of DNA from Neanderthals to humans. By scanning the 1000 Genomes dataset, the authors find the selected haplotype occurs at different frequencies among diverse populations.

Overall I found the statistical analysis presented in the manuscript that links the COVID-19 risk haplotype with the Neanderthal genome convincing. However, without further replication of the result in the case/control data, and speculation of how this adaptation is related to COVID-19 or an ancient pathogen, it is hard to assess its impact on understanding the disease or on the clinical outcomes.

Specifically, I would suggest the following analysis:

1. UKBiobank has released exome sequencing data and the COVID-19 statuses of its participants. The authors can utilise this data to further characterize the observed adaptation within severe

COVID-19 cases.

2. To check whether the COVID-19 risk locus is under strong local selection, the authors can compute the fixation index (F_{st}) between all populations from the 1000 Genomes Project. This analysis can help readers to have a genome-wide view to see if variants in this region are more differentiated than one would expect given the average genome-wide differentiation.

3. It will be especially useful if the authors can add comparisons and discussions of the COVID-19 risk locus to other well-established examples of human adaptation (e.g. EPAS1, SLCA45A2).

Author Rebuttals to Initial Comments:

Referees' comments:

Referee #1 (Remarks to the Author):

Report on 'The major genetic risk factor for severe COVID-19 is inherited from Neandertals'

This manuscript investigates the possibility that a region on chromosome 3 that was recently associated with risk for severe Covid-19 outcome has been inherited from Neandertals. The authors compare the specific variants in the associated region with different Neandertal genomes as well as genetic data from several present day human populations. They find that many variants that are in strong linkage to the top Covid-19-associated variant are also found (in homozygosity) in a Neandertal genome from the Vindija cave, a pattern also supported by further Neandertal genomes. The authors conclude that this haplotype was indeed inherited from Neandertals, and support their conclusion with simulations of shared ancestry and by phylogenetic analysis of the haplotypic region across modern human populations.

The analysis is obviously very timely, and a better understanding of the genetic risk variants and their distribution is of very broad interest to the scientific community, but also of particular importance for any public health efforts in the face of this pandemic. The evolutionary perspective taken here is particularly intriguing.

The hypothesis is straight-forward and plausible. Inheritance of Neandertal DNA in modern humans is well established, and this might be a particular interesting case. The overall analysis and presentation of the data and results appears sound and I can identify no major flaws, but there is some information missing that would improve the accessibility and interpretability of the results. I have detailed these points below.

I am a bit concerned about the use of the Covid19hg data. I know that this data is made available for immediate research use, but I am not sure whether the publication of this data is also 'free', before it has been published by the Covid19hg initiative. The section on the covid19hg website about 'Terms of use of results and authorship' is still empty, stating only a note 'Coming soon'. So I would urge the authors to obtain a letter from the coordinators of Covid19hg stating that they approve the publication of these results, or provide other verifiable proof that the publication of these results does not represent misconduct.

Letter from Andrea Ganna of the Covid19HG Consortium is attached. We note that HZ was an early member of the Consortium, and both HZ and SP are contributing data to the Consortium. We also note that the manuscript was shared with the coordinators of the Consortium before it was put on bioRxiv and submitted to Nature.

I would find it important to see the general frequency of Vindija 33.19 Neandertal ancestry along the chromosome 3 in Europeans (this is where most subjects in this data are from), in order to judge the pattern observed in the Covid-19-associated region. If the frequency of Neandertal ancestry is generally low, this would highlight the particular pattern observed for the Covid-19-associated locus.

Please, see figure below from Steinrücken *et al.* (2018) (Molecular Ecology 27:3873-3888). The introgressed fragment is visible at ~46 Mb (arrow added). Having reviewed the literature, we find introgression at this locus in all introgression maps published to date (see new Supplementary Table S2, summarizing previous data).

[redacted]

At the end of the first paragraph (main text) the authors state that the association with the ABO locus was not replicated by the Covid19hg data. However, this might simply be due to the fact that the primary endpoints of the two studies differ ('severe respiratory failure' vs. 'hospitalization'). So I suggest adding something like the following note to this sentence: ", possibly due to different case definitions between the studies". This will not lower the authors' point that the chr3 locus is the major association with Covid-19, but it will prevent the notion that the earlier study might have yielded erroneous results.

We have added the suggested sentence.

I do not agree that positive selection at this region of interest is unlikely. Clearly, the genes in this region (which include cell-surface receptors that are often targeted by viruses to enter host cells) have more functions/phenotypes than just the effect on Covid-19 risk, so it is certainly possible that variants in this region were selected for other reasons, but now one (or more) of the variants yield a disadvantage in the face of Covid-19. The variants under positive selection and those now associated with Covid-19 could also be different and just be linked. Such effects of genetic draft are well described for the HLA region (Lenz *et al.* 2016 MBE, doi:10.1093/molbev/msw127). In fact, this would be a good explanation why this haplotype was maintained in modern humans in the first place, so I would urge the authors to include this as a possibility, especially at the end of the discussion, where they speculate about the role of this haplotype for the health of Neandertals.

The reviewer makes a valid point. In fact, work from the Akey lab have shown selection of this Neandertal haplotype in Bangladesh (Browning *et al.* (2018) Cell 173:53–61). We have now added a paragraph that discusses selection on page 4 (paragraph 2).

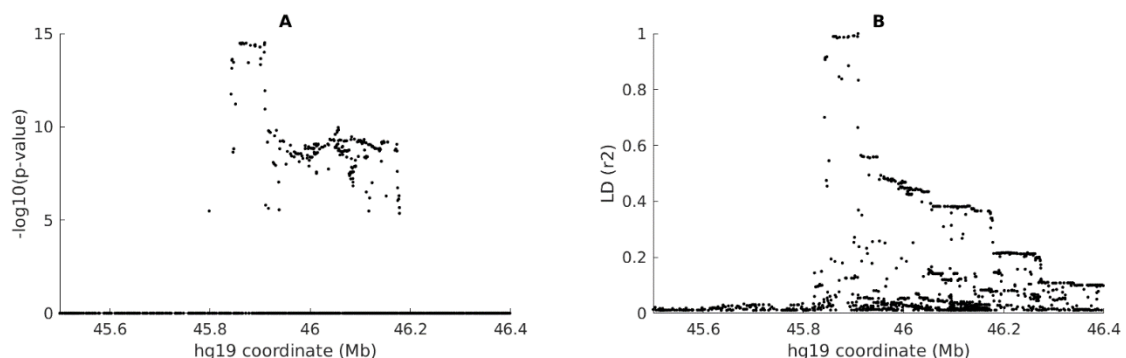
Along those lines, it should also be stated more clearly in the text, that the model-based rejection of the possibility for this haplotype to derive from a common ancestor is based on the assumption of neutral evolution (or so I assume). I guess that a model with balancing selection on this haplotype would yield a different result.

Our choice of requiring $r^2 > 0.98$ to define a haplotype is very conservative. In fact, as can be seen in Fig. 1B, a substantial number of individuals carry longer contiguous haplotypes of

Neandertal-like sequences, reaching a length of ~400 kb. In addition to the calculation for the 49.4 kb core haplotypes, we now make the calculation also for the longer haplotypes of 333.8 kb (Suppl. Fig. S1). As expected, it is even more strongly rejected that this haplotype is derived from the common ancestor ($p=10^{-26}$). We note that the previously published introgression maps, using different methods and different aspects of the data, all identify introgression from Neandertals at this locus (new Suppl. Table S2).

The shared red color between Fig1A and 1B suggests similar effects, but Fig1B does not actually show association with Covid-19, it only shows linkage to the lead SNP in Fig.1A. I find this somewhat misleading and thus suggest to use a different color to indicate LD patterns in Fig1B.

We have removed the red color from Fig. 1A to avoid confusion and indicated the core haplotype with a bar in Fig. 1B in the main text. It should be noted that the association correlates almost perfectly with LD to the lead SNP (rs35044562), see figures below (A: Association to “hospitalized COVID-19”; B: LD to lead SNP). Thus, there are no significant associations to the disease ($p < 5 \times 10^{-8}$) which are not also in LD with the lead SNP.



The fact that the association signal extends (at a lower level) to 400-500 kb is brushed over with a short sentence in the legend of figure 1, stating that “some individuals carry a Neandertal-like haplotype that extends to ~400 kb”. It is not stated how this information was obtained nor how it was verified. I think this needs to be explained in more detail, as it suggests that it is not only one haplotype that drives the association with Covid-19, as could be understood from the description of ‘the Neandertal haplotype’ throughout the text. I guess this signal comes from recombination/haplotype decay, which is in agreement with the scenario laid out here, but it should be described more clearly.

We have clarified this in the manuscript (page 2, paragraph 2). The association signal is driven by the 50 kb “core Neandertal haplotype” as shown in the figure above. Some individuals carry longer haplotypes, extending up to ~400 kb.

In the main text, the authors mainly focus their analyses on the SNPs that are in very high LD ($R^2 > 0.99$) with the top SNP from Ellinghaus et al. 2020. However, in Fig1B, to which the main text refers for these SNPs, all SNPs with an $R^2 > 0.1$ are marked in the same color (red), creating the

impression that all red SNPs are in very high LD. I find this somewhat misleading and think that it would be helpful if these top SNPs (with $Rsq > 0.99$) were colored differently in Fig. 1B, so that information is conveyed more transparently.

We have now indicated the core haplotype SNPs in Fig. 1B with a black bar. We have now also changed all analyses and figures to use the lead SNP from the Covid19HG Consortium (rs35044562) rather than the Ellinghaus *et al.* lead SNP to avoid any confusion.

In Fig.2, it is not clear why there are multiple haplotypes in the shaded box. In the text it is stated that these haplotypes are associated with Covid-19, but it is not clear how this was determined. Do they all carry the top associated variant? This is a bit confusing as otherwise in the text mainly one haplotype is mentioned (see also Fig. 3). Please clarify.

All of the haplotypes that carry the lead SNP are marked by the shaded box. This is now clarified in the figure legend.

Also, how robust is this clustering result with regard to the chosen substitution model, e.g. how would it look like with Tamura-Nei model?

We have now estimated the tree also with the Tamura-Nei model. Its structure is very similar to the one presented (see below).

[redacted]

In Fig.3, the frequency of the Neandertal haplotype is shown across 1KG populations. However, it is not clear how the data and the populations were selected for this analysis (the 1KG data contains more populations than what is shown here). This analysis needs to be described in more detail. I also find the choice of symbols for the populations where the haplotyp is missing suboptimal, as they give much more weight to 'missingness' than to very low frequency.

All 1KG populations are in fact used in the figure. We have modified the figure according to the suggestions, now using pie charts to more clearly visualize the allele frequencies.

What is 'the Neandertal risk variant', described in the population frequency analysis? Is it the top-associated variant from Ellinghaus *et al.* 2020 or the one described in Fig. S1 as being in perfect LD with that top SNP or something entirely different. This should be more clearly described in the text.

The lead SNP in the newly released GWAS is rs35044562 where Neandertals carry the derived allele. The Ellinghaus *et al.* lead variant is in LD with this variant in Euroasian populations. To avoid confusion, we now use the lead SNP from the Covid19hg Consortium throughout the manuscript. We apologize that this was not clear in the manuscript.

Referee #2 (Remarks to the Author):

This study from Zeberg and Paabo tests the hypothesis that major genetic risk factor for severe COVID-19 on chromosome 3 was introduced by Neanderthal introgression into modern humans. Given that relatively little is known about the host genetics associated with COVID-19, this is a potentially exciting result that has major implications for both our understanding the host genetic response to COVID-19 and to the genetic impact of Neanderthal introgression.

The main lines of evidence offered to support a Neanderthal origin for the risk haplotype on chromosome 3 (following techniques used by Huerta-Sanchez and colleagues in their article to show that the variant at EPAS1 that conferred adaptation to high-altitudes was introduced by Denisovan introgression) are that:

- 1) the risk haplotype for COVID-19 is genetically closer to the Neanderthal haplotypes (specifically the Vindija Neanderthal) relative to other haplotypes segregating in modern humans
- 2) the risk haplotype is too long to have been an ancestral polymorphism that was inherited unrecombined across human populations, and
- 3) the geographic distribution of the risk haplotype wherein the risk haplotype is at appreciable frequencies in south Asia, at lower frequencies in Europe but is absent in Africa.

While the first line of evidence is convincing, the second and third lines are less so.

1. The authors argue that the Neanderthal risk haplotype is almost completely absent in Africa. This statement contradicts Ellinghaus et al. 2020 (Figure S8) that shows substantial frequencies of the risk allele in African populations with frequencies ranging from ~2% in Sierra Leone to ~7% in Nigeria and Kenya (comparable to the frequencies seen in Europe and higher than those in East Asia).

Thus, the authors need to clarify this contradiction.

The lead SNP in the newly released GWAS from the Covid19HG Consortium is rs35044562, which is virtually absent in Africa.

In the 1000 Genomes Project, the lead SNP from the Covid19HG Consortium and the lead variant from Ellinghaus *et al.* are in LD ($r^2=0.99$) in Eurasian populations. The Ellinghaus *et al.* SNP does exist in Africa at a frequency of 5.3%. However, in Africa it is not associated with other Neandertal-like variants. It thus probably exists in Africa either as the result of a parallel mutation, or (more likely) as a result of incomplete lineage sorting from the common ancestor of Neandertals and modern humans.

However, in the larger GWAS from the Covid19HG Consortium, the p-value for the Ellinghaus *et al.* lead SNP is $p=3.6e-14$ whereas the p-value for rs35044562 (which is unique to the Neandertal haplotype) is 10-fold lower, *i.e.*, $p=3.1e-15$.

To avoid confusion, we now use the lead variant from the Covid19HG Consortium (which is more strongly associated with severe COVID-19) throughout the manuscript. This was not clear in our original manuscript. We thank the reviewer for raising this point.

It is plausible, however, that an introgressed haplotype could have attained high enough frequencies in Africa through gene flow into Africa from non-African populations. To show that the risk haplotype is indeed introgressed, the authors should inspect recent maps of introgression (e.g. Chen et al. 2020). Such an analysis could further clarify the evolutionary history of the risk haplotype in modern human populations.

See answer above. The lead variant from the Covid19HG Consortium does not exist in Africa.

We have added Suppl. Table S2 that lists previous genome-wide maps of Neandertal introgression, which have identified this locus as introgressed. These introgression maps rely on different methods, different recombination maps, and in some cases no recombination maps.

2. An analysis of the length of the haplotype is used to argue against the possibility of incomplete lineage sorting producing the patterns we see today. This analysis uses a recombination of 0.53 cM/Mb.

What recombination map was used in this analysis? Given the uncertainties of recombination maps at kilobase scales, it would be essential to redo this analysis using either admixture-based and/or pedigree-based maps.

The recombination map used is pedigree-based (Kong *et al.* (2002) Nature Genetics 31:241-247). This reference has now been added also where the calculation is mentioned in the text.

We have also included references to seven previous studies (page 3, paragraph 1) that have identified introgression from Neandertals at this locus using different recombination maps and different methodological approaches (see new Suppl. Table S2). Several of these studies use methods independent of recombination maps (S*, Skov et al., 2020).

Referee #3 (Remarks to the Author):

The authors find the recently identified COVID-19 risk locus in Chromosome 3 has a haplotype structure that can be explained by introgression of DNA from Neanderthals to humans. By scanning the 1000 Genomes dataset, the authors find the selected haplotype occurs at different frequencies among diverse populations.

Overall I found the statistical analysis presented in the manuscript that links the COVID-19 risk haplotype with the Neanderthal genome convincing. However, without further replication of the result in the case/control data, and speculation of how this adaptation is related to COVID-19 or an ancient pathogen, it is hard to assess its impact on understanding the disease or on the clinical outcomes.

Specifically, I would suggest the following analysis:

1. UKBiobank has released exome sequencing data and the COVID-19 statuses of its participants. The authors can utilise this data to further characterize the observed adaptation within severe COVID-19 cases.

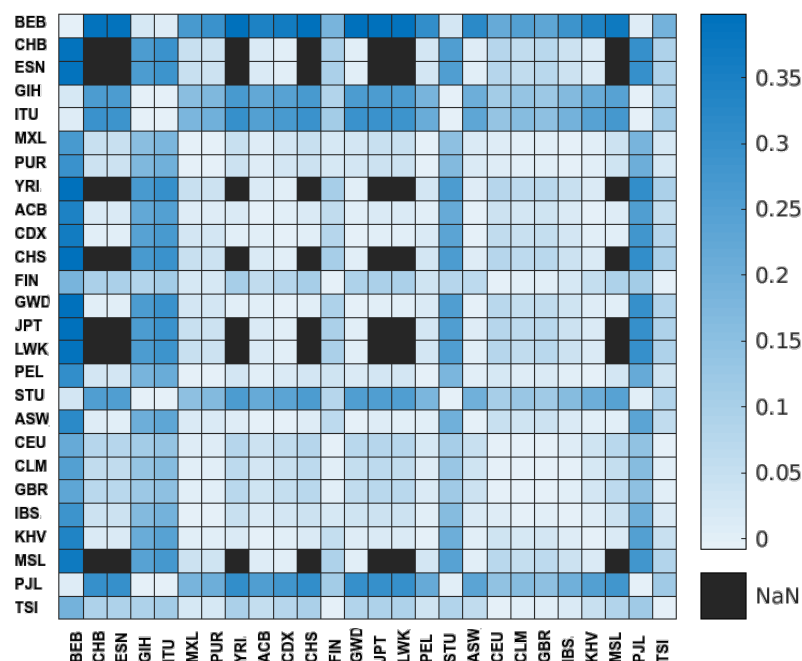
The meta-analysis presented rely on several data sets, including the UKBiobank. Fortunately, a SNP on the core Neandertal haplotype is included on the UKBB array (rs71325088). The results of the UKBB and the other studies are shown below. We do not have access to the UKBB exome data.

COVID-19 Host Genetics Initiative, <https://www.covid19hg.org/>

[redacted]

2. To check whether the COVID-19 risk locus is under strong local selection, the authors can compute the fixation index (F_{st}) between all populations from the 1000 Genomes Project. This analysis can help readers to have a genome-wide view to see if variants in this region are more differentiated than one would expect given the average genome-wide differentiation.

We thank the reviewer for raising this point that was also raised by reviewer 1. Since the submission of the manuscript, we have become aware that Browning *et al.* (2018) (Cell 173:53–61) have identified this locus as positively selected. Given the careful analyses performed by them we prefer to refer to their work on page 4, paragraph 2, where we now point out that there is previous evidence that this Neandertal haplotype was positively selected in Bangladesh. However, for information, F_{st} values between all 1000 Genomes populations are shown here:



Three letter codes denote 1000 Genomes populations. Note the high F_{st} values for BEB (Bangladesh) vs other populations, compatible with the analysis by Browning *et al.* (2018, Cell, 173:53–61).

It should be noted, however, that F_{st} values are hard to interpret for introgressed loci given that gene flow from Neandertals is geographically localized, perhaps even in ways that are not yet understood. The analysis performed by Browning *et al.* (2018) is therefore more stringent.

3. It will be especially useful if the authors can add comparisons and discussions of the COVID-19 risk locus to other well-established examples of human adaptation (e.g. EPAS1, SLCA45A2).

We have now added references to the examples suggested by the reviewer (page 4, paragraph 2), as well as the toll-like receptor locus (Danneman *et al.* (2015) *Am J Hum Genet*, 98:22-33) and the progesterone receptor (Zeberg *et al.* (2020) *Molecular Biology and Evolution*, msaa119).

Reviewer Reports on the First Revision:

Referee #1 (Remarks to the Author):

I have already reviewed the initial version of this manuscript and I appreciate the authors' effort in addressing my and the other reviewers' comments. I think that the manuscript has improved in clarity and substance, and I have no further concerns or comments. I congratulate the authors on this fine work.

Referee #2 (Remarks to the Author):

The authors have addressed my concerns. I am happy to support publication of this study.

Referee #3 (Remarks to the Author):

Thank you to the authors for this quick response. I'm afraid that some of their answers did not fully address my previous comments:

1. For my second point regarding F_{st} , I was hoping to see a genome-wide F_{st} versus max freq difference (similar to Huerta-Sánchez *et al.* 2014 Fig1), highlighting genes that harbour SNPs previously identified as having strong local adaptation. This will provide readers a global view of the strength of this covid-19 risk locus compared to others (e.g. EPAS1).

2. In addressing Reviewer #2's comment, the authors noted the different LD structures are different in the region among continental populations, and argue the reason for choosing the lead variant (rs35044562) to look up was due to a lower p-value observed in the covid19-hg consortium. However, it is not a convincing argument just to look at the p-values to determine which variant is more likely to be 'causal'. Without a fine-mapping study in the region, authors should at least check whether the reported lead variant (rs11385942) from Ellinghaus *et al.* is also in line with their presented findings to give a second line of evidence for suggested introgression. If they are in the same risk haplotype, I shall expect they reach the same conclusions.

Author Rebuttals to First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

I have already reviewed the initial version of this manuscript and I appreciate the authors' effort in addressing my and the other reviewers' comments. I think that the manuscript has improved in clarity and substance, and I have no further concerns or comments. I congratulate the authors on this fine work.

Referee #2 (Remarks to the Author):

The authors have addressed my concerns. I am happy to support publication of this study.

Referee #3 (Remarks to the Author):

Thank you to the authors for this quick response. I'm afraid that some of their answers did not fully address my previous comments:

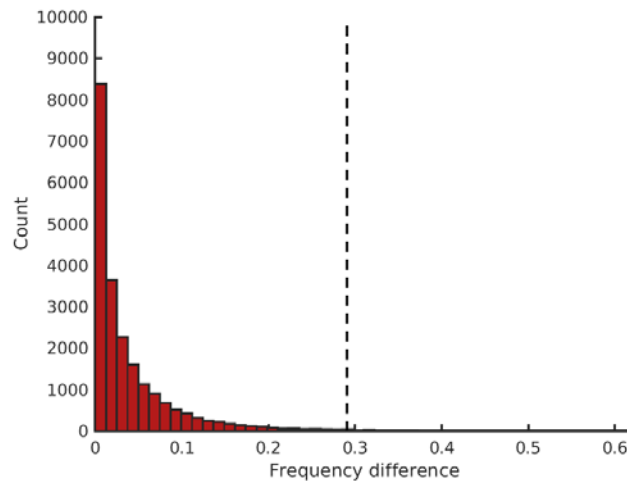
*1. For my second point regarding F_{st} , I was hoping to see a genome-wide F_{st} versus max freq difference (similar to Huerta-Sánchez *et al.* 2014 Fig1), highlighting genes that harbour SNPs previously identified as having strong local adaption. This will provide readers a global view of the strength of this covid-19 risk locus compared to others (e.g. EPAS1).*

We assume that “the strength of this covid-19 risk locus” implies the strength of selection to which this locus may have been subjected in past modern human populations relative to other loci such as EPAS1 in Tibet. However, such direct comparisons are difficult between population pairs that have different demographic histories because differences in patterns of variation will not only be affected but differences in putative past selection regimes but by differences in demographic histories. For example, smaller populations will have larger allele frequency differences simply because they are of smaller size.

The maximum allele frequency difference between any two populations in Fig 1. in Huerta-Sánchez *et al.* (2014) (*Nature* 512:194-197) is ~50% or higher. In our case, the allele frequency difference for the chromosome 3 COVID-19 risk locus has the biggest frequency difference between South and East Asians. These are vastly bigger populations than Tibetans making large allele frequency differences less likely. The maximum frequency in South Asians is 37% in Bangladesh. It would thus not appear in the analysis of Huerta-Sánchez *et al.*

To put the putative past selection at the COVID-19 risk locus into a relevant perspective, we now compare it to other Neandertal haplotypes in Asia. To do this, we computed the frequency differences between South and East Asia for all introgressed Neandertal haplotypes identified by Chen *et al.* (2020) (*Cell* 180:677-687) that occur in South and East Asia. We find that the frequency difference of the chromosome 3 risk haplotype is unlikely to be observed by chance ($p = 0.006$). See figure below (also shown as new Extended Data Figure 6).

This is in agreement with the previously claims of selection of this haplotype in South Asia (Gittelmann *et al.* (2016) *Current Biology* 26:3375-3382, Browning *et al.* (2018) *Cell* 173:53-61, Browning & Browning (2020) *bioRxiv* 2020.07.15).



Extended Data Figure 5. Frequency differences between South and East Asia for haplotypes introgressed from Neandertals. The dashed line indicates the frequency difference for the Neandertal haplotype conferring risk to severe COVID-19.

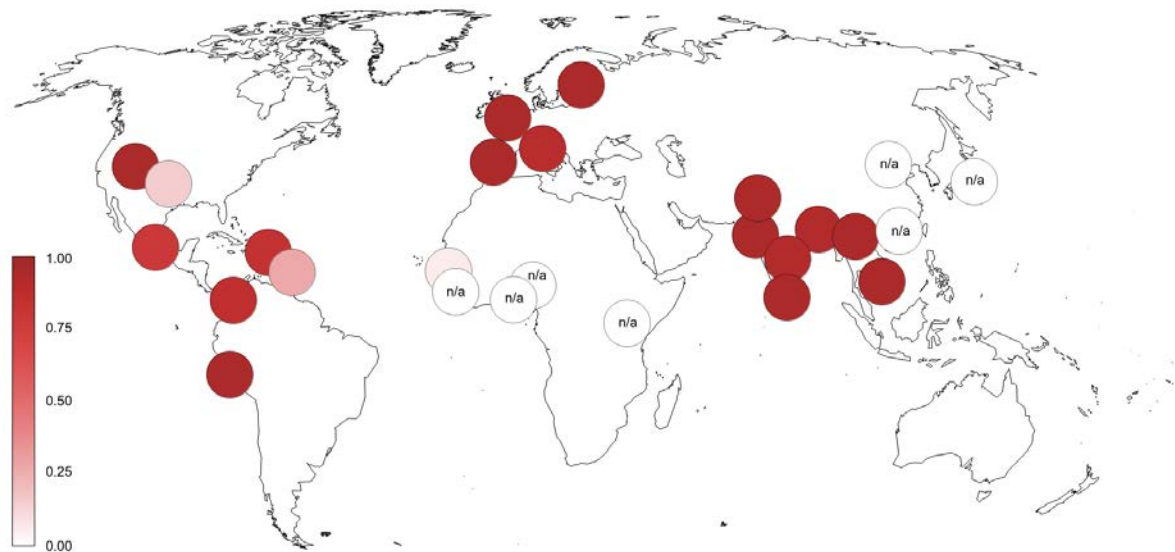
The unusual frequency difference is compatible with positive selection in South Asia, negative selection in East Asia, and/or a combination of the two. Perhaps this selection has been of an infectious nature. For example, previous outbreaks of corona viruses in East Asia is a possible reason for the almost complete absence of the haplotype in the latter region. However, that is completely speculative at this point.

We have now modified our manuscript to incorporate this additional analysis (page 4, third paragraph, Extended Data Figure 5).

2. In addressing Reviewer #2's comment, the authors noted the different LD structures are different in the region among continental populations, and argue the reason for choosing the lead variant (rs35044562) to look up was due to a lower p-value observed in the covid19-hg consortium. However, it is not a convincing argument just to look at the p-values to determine which variant is more likely to be 'causal'. Without a fine-mapping study in the region, authors should at least check whether the reported lead variant (rs11385942) from Ellinghaus et al. is also in line with their presented findings to give a second line of evidence for suggested introgression. If they are in the same risk haplotype, I shall expect they reach the same conclusions.

Responding to this concern by Reviewer #3, we make two points:

- i) We show that the Ellinghaus *et al.* variant and the variant of the COVID-19 Host Genetics Initiative are in linkage disequilibrium in Eurasia ($r^2 = 0.99$ in the 1000 Genomes Project) in more detail than we did previously in the response to Reviewer #2. Thus, the same Neandertal haplotype is tagged by both variants in Eurasia and thus conclusions drawn are the same.
- ii) Despite the high LD between the two variants, the Ellinghaus *et al.* variant, which also exists in Africa, is less likely to be the causal variant as indicated by the statistical fine-mapping of the locus performed by the COVID-19 HGI.



i) Below, we show a map of linkage disequilibrium between rs35044562 and rs11385942 in all 26 populations of the 1000 Genomes Project.

Extended Data Figure 3. Linkage disequilibrium between the Ellinghaus *et al.* index variant (rs11385942) and the index variant of the COVID-19 Host Genetics Initiative (rs35044562). Shades of red indicate the extent of linkage disequilibrium (r^2) in the 1000 Genomes populations. Populations labelled with "n/a" are monomorphic for the protective allele of rs35044562. The Ellinghaus *et al.* index variant does not have any genetic variants in LD ($r^2 > 0.8$) in African populations.

Of the **410** chromosomes that carrying the risk allele of the lead variant from the COVID-19 HGI among the world-wide 1000 Genomes individuals, **408** (99.5%) also carry the risk allele of the Ellinghaus *et al.* lead variant. The probability of observing this level of co-segregation by chance is $p = 7 \times 10^{-524}$.

Note also that the risk haplotype identified by Ellinghaus *et al.* is the same haplotype as the one we identify as the "core" haplotype originating from Neandertals. See Fig. 3A from their paper below.

Ellinghaus *et al.*, Genomewide Association Study of Severe Covid-19 with Respiratory Failure,
<https://www.nejm.org/doi/full/10.1056/NEJMoa2020283>

[redacted]

Fig. 3A from Ellinghaus et al. (left) compared to Fig. 1B in our manuscript (right). Red circles (added by us) indicate the core Neandertal haplotype.

The Ellinghaus *et al.* index variant occurs also in Africa whereas the COVID-19 HGI index variant is absent in Africa. One question is then if the occurrence of this Ellinghaus *et al.* variant in Africa is due to an independent mutation in Africa or is inherited from the common ancestor of Neandertals and modern humans. Assuming an effective population size of 7,500 for Africans (Wang *et al.* 2016 *Heredity* 117:193–206) and a mutation rate of 2.5×10^{-8} per site/generation, the expected rate of observing a new mutations at any one nucleotide positions among 661 African individuals is 0.005. Hence, the Ellinghaus *et al.* lead variant is therefore unlikely to be the result of an independent mutation in Africans. Rather, it was present in the common ancestor to Neandertals and modern

humans. This is further supported by the fact that in Africa the haplotype is so short that there are no other genetic variants in LD ($r^2 > 0.8$), indicating that the mutation is of old age. Thus, the risk haplotype identified by the two association studies is not present in Africa.

Please, note that our first submission of this manuscript used the Ellinghaus *et al.* index variant and reached the same conclusions as the current manuscript. In the resubmission we switched to the COVID-19 HGI lead variant to avoid confusion as this is likely to be the variant used in future work (see also below).

ii) A formal possibility is that the Ellinghaus *et al.* index variant would be causative and the Neandertal haplotype would 'only' be responsible for transferring this ancient variant to non-African populations, where 97% of the deaths have occurred and thus association studies have been performed (ECDC, 2020.08.16). The Neandertal haplotype would still "be to blame", but the causal variant would originate in the common ancestor of modern humans and Neandertals. This is, however, less likely because a fine-mapping performed by Kanai & Finucane for the COVID19 HGI (<https://www.covid19hg.org/>; *In silico follow-up results: Fine Mapping*) find that the Ellinghaus *et al.* variant is not included in the 95% credible set for being causative. Below, we show their results. However, in our manuscript, we refrain from speculating about what may be the causal variant. Our work makes the simple point that the haplotype associated with increased risk for severe COVID19 in Eurasia is of Neandertal origin. What feature in this haplotype causes this is the subject of ongoing work by many groups.

COVID-19 Host Genetics Initiative, <https://www.covid19hg.org/>

[redacted]

Statistical fine-mapping performed by the Covid-19 Host Genetics Initiative. The y-axis gives the posterior inclusion probabilities ("PIP") and the x-axis chr3 coordinates. The probability of rs35044562 being causative is substantially higher than rs11385942. Arrow and rsIDs for index variants added by us.

Reviewer Reports on the Second Revision:

Referee #3 (Remarks to the Author):

I thank the authors for their thorough replies in addressing my comments. I have no additional comments, and happy for the publication of this exciting finding.

Author Rebuttals to Second Revision:

Referee #3 (Remarks to the Author):

I thank the authors for their thorough replies in addressing my comments. I have no additional comments, and happy for the publication of this exciting finding.

We thank the three referees for their comments which have strengthened the manuscript