nature portfolio

Peer Review File

Rare disease gene association discovery in the 100,000 Genomes Project

Corresponding Author: Professor Damian Smedley

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The manuscript presents the results from applying a gene burden analytical framework to discover rare disease-gene associations. Although the approach applied in the manuscript is not novel (see their reference 4), my understanding is that its extension, the development of an R package, its application to a large dataset of more than 35K cases and relatives, and the new associations presented are novel.

Comments:

L117-L122: consider if references 1 (Xiao et al) and 2 (Halley et al) are the most appropriate to indicate prevalence of RDs in the UK and the % of molecularly undiagnosed cases despite genome sequencing. For the latter, wouldn't it be more appropriate to directly refer to the diagnosis yield from 100K?

L133-L141: unclear what is being presented in THIS manuscript for the first time (and is therefore novel) and what was previously developed and already published unless reading the referenced material (example: is the R framework completely novel or just its "extension"?)

L155: authors claim that cases and controls could be defined in other ways, such as "by specific phenotypic annotations or phenotypic clustering". Have the authors tested this? If so, how different are the results compared with those from the approach they present in the manuscript? When is one approach more appropriate than another? How can a potential user decide?

L171, Fig 1 and Online Methods: it took me a while to figure out the "pipeline", since L171 and Fig1 start with 35,008 families, while Online methods start with 35,548. Might be useful to try to mention in the main text, or at least in Fig 1 legend, that this is a "filtered" cohort and refer to Online methods.

L198-L198 (and other parts of the manuscript): consider discussing how some cases that might have previously received an incorrect or incomplete diagnosis (e.g. for a digenic situation) could be be limiting the power of the approach.

L205 - I am struggling to understand where the "87" comes from.

L215-L218 - could these enrichments indicate these associations might be false positives?

L415-L416 - unclear; consider splitting in two sentences.

L466 - it would be useful to have a description of the efforts done (and potential findings) through GeneMatcher. Have they considered other gene matching platforms or MatchMaker Exchange (which would give them access to larger numbers)?

L474 - given the assumption/limitation indicated here, would have it been more appropriate/better to define cases and controls "by specific phenotypic annotations or phenotypic clustering"? (see comment above).

L603-L607 - unclear what is the rationale supporting the indicated limitations to "maintain statistical validity and power".

L639-L644 - seems gramatically incorrect. Also, a paranthesis seems to be missing (or there is one too many).

L646: "[...] or strong (12-18 points)" what? (missing a word here?)

L646-L648: does not seem to be grammatically correct.

- Although the authors mention a study using BeviMed on approximately the same cohort, they have not compared their approach to other ones or discuss why theirs is better / more appropriate / Pros_Vs_Cons in comparison to other methods. A direct comparison with other methods would be useful but, if this is not possible or unjustified, some information for the reader to decide when/why to choose the authors approach would be appreciated.

- Can the framework be used (or easily adapted to be used) on federated datasets in which it is not necessary to have access to the full set of variants from each dataset? This would make it possible to increase the sample size without having to pool data centrally, overcoming many ethical, legal and technical challenges.

- It would be useful to define "proband" and "case".

- Fig 1: see comment above.

- Fig 2 unreadable in printed form. Consider splitting in more than one page or including only the blue (88 triaged novel signals) in this Figure and add the rest as supplementary.

- Figs 3 and 4: unreadable in printed form. Try to improve readability and/or consider publishing only one or two of the main panels as a Figure and the rest as supplementary figures. Legends could be more detailed (ex. missing info about what is each subpanel and how to interpret it)

- Supp Fig 1: "highly" should be "high"? Also, abbreviations missing (BP, CC, MF).

The software developed and the outcomes of this manuscript are of interest to people in the fields of human genetics and bioinformatics.

The code of the open-source R framework (geneBurdenRD) and Exomiser are both available online. The installation and usage of geneBurdenRD seem to be well documented, although I have not tried to install and run it.

To the best of my knowledge, the manuscript does not have any flaws which prohibit its publication, the conclusions are original and the statistical tests seem appropriate, although I have indicated in my report where additional details might be useful for the reader.

Referee #2

(Remarks to the Author)
In this manuscript, the authors applied a gene burden analysis approach to a very large cohort of rare disease patients from the 100,000 Genomes Project and identified 88 novel gene-disease associations, in addition to identifying 190 previously known disease associations. They evaluate the evidence for each one according ClinGen's gene-disease validity framework, 38 of which meet moderate or strong levels of evidence with another 50 at limited. This demonstrates the utility of a large rare disease research datasets for delivering novel gene discoveries and candidates. Over all, this paper is a very nice body of work with major contributions to rare disease gene discovery. I have a few areas of feedback that, if addressed, could improve the paper.

There does not appear to be any list of the full set of variants that implicate the novel 88 gene-disease associations. This would be important to include so that the evidence is clear and readers and reviewers can better judge the full range of evidence for each novel gene-disease relationship and proposed mechanism of pathogenicity.

It's also not clear if each variant was separately classified according to pathogenicity versus only considering the total gene burden. I would recommend individual variant classification for those variants in the genes meeting strong or moderate ClinGen evidence. It would be useful to know how many individual variants reached pathogenic or likely pathogenic according to current ACMG/AMP standards versus how many were VUS on their own. I imagine some may have had sufficient evidence for reach P/LP based on predicted LOF impact, de novo occurrence, being in trans with another P/LP variants, etc. It's worth noting that ClinGen guidance suggests that once a gene-disease relationship meets strong or definitive, variants can be classified up to pathogenic, but if only moderate, the ceiling is LP and if limited, the ceiling is VUS. The variant classifications and supporting evidence should also be submitted to ClinVar so these variants are findable by the community.

The authors use the terminology "a ClinVar known pathogenic". The authors have not defined what that description means

and I would hope that it's not defined by "one or more submitters to ClinVar called it pathogenic" as one cannot blindly trust all submissions to ClinVar as accurate. Instead, I would encourage the authors to classify variants according to the ACMG/AMP classification standards. Evidence and rationales submitted to ClinVar can be useful to contribute to variant classification, but an independent review of all evidence and entries in ClinVar is recommended.

The authors list "1 or 2 LOF" each for DYSF, KCNQ1, RYR1 and COL4A3 which have primarily missense variants, and in some cases, LOF is known to cause a distinct disease. Have the authors rigorously assessed the likelihood of rescue for these pLOF variants? A framework to guide this can be found in PMID: 37633279. Similarly, the pLOF variants found in gnomAD for UNC13A could be assessed with this framework.

The authors note that de novo annotations were restricted to "only trios or larger families where de novo calling was possible and provided by the user". This seems like a missed opportunity to not do systematic de novo assessment on the call set for all trios. At a minimum, it would be useful to know how much of the data had de novo calling applied. I was particularly surprised to see that de novo variants were not observed in any of the novel associations. It would be important to note if this is due to only a small portion of the cohort having trios and de novo calling vs some distinction in the characteristics of the diseases for the novel gene-disease associations (age of onset, inheritance patterns, penetrance, etc).

The authors noted "Potential associations were further filtered to 88/316 (28%) by removing those in which (i) the gene was a non-protein coding RNA one (31/316)". It would be useful to understand why non-protein coding RNAs were excluded. This seems like a missed opportunity for discovery.

For the 88 novel gene-disease associations, the authors note that they applied ClinGen's framework for gene-disease validity. This is excellent but I had a few questions/suggestions:
A) The section of the methods describing the process is labeled "Second clinical expert (GeCIP) and in silico triage" which does not really describe the gene-disease validity assessment. I would suggest labeling this section "Application of ClinGen framework for gene-disease validity" or something like that. Also, the first sentence notes that its automated, yet nothing described sounds automated! I will also note that its not clear how this step relates to the triage in general. Please clarify.

B) Was the evidence from the "5 interim confirmations" included in the gene-disease validity assessment or not? Just because there is a publication, does not mean the gene-disease association is now established so I would suggest you either include the published evidence in your primary assessment, or if not, then perform the assessment a second time with the published evidence included and make it clear what the current classification is. I might also refer to these published claims as "published gene-disease associations" or "published support" instead of "confirmations" which implies the mere fact of publishing something makes it definitive.

C) It's also not clear what attributes placed 7 genes into the "compelling candidate" compared to other genes (subset of moderate??).

D) On a related note, the authors state "Of the remaining 83 novel associations, nine had prior functional data fitting the likely disease mechanism and a high ClinGen classification score (>= 8)". Its not clear what "high" means here as this includes most of the spectrum of ClinGen's moderate evidence (7-11 points). I might suggest continued use of ClinGen's terms or using the score alone without the word "high" if you want to remain using 8 and above.

Under the section of the paper describing COL4A3, the authors note that eleven of the cases already had PKD1 (or 2 with PKD2) diagnoses and raise the possibility of digenic inheritance or modifiers. After a quick look at the abstract of the referenced paper (citation 43) there may be some plausibility to this mechanism for COL4A3/COL4A4/COL4A5, given the heterotrimer configuration of these proteins. However, it's not clear if that is true for PKD1 and PKD2. Its also worth noting that PKD1 is one of the most polymorphic genes in the human genome with 1.8 million missense allele counts in the general population in gnomADv4 (average missense burden being 2.3 missense variants per person). Only 26 of the 5932 unique missense variants in gnomAD are classified as P/LP. As such, I would strongly suggest reviewing the detailed evidence on each variant before concluding digenic inheritance for any of these cases. None of the variants are shared for independent review..

The authors note "Further evidence is necessary before all the novel associations described here can be used clinically for diagnostics, counselling, and management." I initially interpreted this to say that none of the findings could be used diagnostically, etc., which I disagree with as some of the gene-disease associations meet strong or moderate, both levels being approved for inclusion in diagnostics panels. However, I'm now guessing the authors meant that not all of them are ready for use (e.g., the limited). Perhaps change "all" to "some" to make this clearer.

In the Methods, the authors describe the selection of coding variants and removal of many other categories such as "coding transcript intron variant". Please clarify if intronic splice site variants (e.g., +/-1,2 and others in the splice consensus sequence) are included or excluded? If included, how far into the intron?

The authors note "Each lolly indicates a variant by its protein change annotated on one single transcript (specified in the plot)…". Please confirm (and change the text to clarify) if these are MANE Select transcripts and if all variation is reported according to MANE Select (or MANE Plus Clinical if a stronger impact is predicted). If not, I'd recommend changing to this now established standard.

Supplemental Table 1: In addition to providing the relative breakdown of cases across the disease groups, top subgroups

and top specific diseases, it would be useful to simply provide the full three lists of disease groups, subgroups and specific diseases and how they are related to each other ontologically. I looked in the prior paper and could not find this articulated there either. This would allow other groups to apply the same categories if they wish.

Please use “variants” instead of “mutations” which is used occasionally in the paper.

Typos:
• KCNQ1 is misspelled as KNCQ1 at least twice.
• ClinVar is mis-capitalized as Clinvar in multiple places.
• In “Supplementary Table 1. Demographics of the 100,000 Genomics Project cohort.”, change “Genomics” to “Genomes”.
• In the sentence “Supplementary Figure 1. Gene ontology enrichment analysis of the GTEx substantia nigra gene co-expression network showing highly enrichment for terms associated with mitochondrial function, a key process in Parkinson’s disease pathophysiology.” “highly enrichment” needs a fix.

Heidi Rehm

Referee #3

(Remarks to the Author)
Key results and originality

In this work, the authors seek novel genes that contribute to rare diseases using gene burden analysis in the Genomics England 100KGP. They claim to have identified 88 novel associations, and highlight in detail a subset of seven that are particularly compelling. This is a very well written and clear manuscript. It addresses a critically important question, which is the need to identify more genetic causes of rare disease. It will be of interest to rare disease researchers and clinicians, and the seven highlighted associations are convincing.

The analytical framework described in this work uses exact tests to compare burden of pathogenic variants defined by Exomiser (Bone et al, 2016) in cases to controls. The framework was previously described (100,000 Genomes Pilot Investigators et al, 2021), and here the code has been adapted for generic application and published along with visualisation scripts.

The 100KGP cohort itself has previously been published (100,000 Genomes Pilot Investigators et al, 2021). The previous publication included gene-based burden testing on ~57K genomes. The present study uses ~73K genomes, so I infer that there is a substantial overlap with those included in the previous publication.

A separate study (Greene et al, 2023) has been published on the same dataset and with a very similar aim (to identify novel disease – gene associations) albeit with a different statistical framework. The authors discuss this in the manuscript.

Given that previous publications have described Exomiser, the analytical framework, and attempts to identify novel disease – gene associations using the same (or a highly overlapping) dataset, in my view the novelty and impact of this study lies primarily in the newly claimed disease-gene associations.

Statistical robustness

I am concerned about the possibility that the novel disease-gene associations the authors claim to have identified could be enriched for false positives for the following reasons:
1) Surprisingly limited overlap with the findings of Greene et al. While as the authors point out, Greene et al use a different statistical framework and different case-control definitions, it is nevertheless a similar framework on the same dataset addressing essentially the same question, so I would expect a high degree of overlap in results.
2) The authors identify 190 known and 316 novel (pre-triage) associations from their analysis. This ratio is surprising. From this kind of analysis, that includes a substantial proportion of diagnosed patients with pathogenic variants in known disease genes, I would expect the majority of significant associations to be known. This suggests that the novel associations may be enriched for false positives.
3) The authors have not attempted to empirically validate the p value threshold they used to define significance.
4) 92% of known associations passed the authors filtering criteria (i) to (iii), but only 28% of novel associations did. This also suggests that the novel associations may be enriched for false positives.

The authors could reassure readers on statistical robustness using the below approaches.
1) Investigating whether the p value distribution of their results is inflated by performing an n-of-1 permutation analysis to generate an appropriate null distribution of p values, and generating a QQ plot.
2) The authors could run an additional ‘negative control’ model using rare variants that are largely expected not to contribute to rare diseases. Eg predicted benign, or synonymous variants. Then compare the p value distribution to the models that use likely pathogenic variants.
3) With frameworks based on exact tests that cannot be corrected for covariates, care needs to be taken to harmonise case-control groups for factors such as age, sex, and ancestry to avoid spurious results. Please could the authors provide some reassurance that their case-control groups have been appropriately harmonised or corrected for relevant potentially confounding factors.

4) How correlated are the studied diseases? Could the authors run a formal pairwise correlation analysis and presented it eg as a heat map, to a) reassure the reader that extensive phenotypic correlation isn't contributing to inflated p values and b) help identify treatment effects.
5) Plot the relationship between family size and number of analysed variants.

Additional data/information

Regarding ancestry:
1) Could the authors please replace self-reported ancestry with genomically-defined ancestry (eg using an algorithm such as Peddy). This would be more complete, accurate, and unbiased.
2) Please include in the results section a description of the ancestral diversity of the cohort. This is currently in a supplementary table and should be more accessible.
3) For the analysis of whether associations are linked to ancestry (page 7), please could full statistics (eg p value) be added.
4) Some of the associations have cases enriched for individuals of non-European ancestry. Can the authors please add some reassurance that those variants are enriched because of the disease and not because of the population structure?

Please add a supplementary table detailing the 226 studied diseases including case and control numbers, and higher level categories.

Could the authors please explain how the threshold of 33.3% was selected, for the triaging filter (iii). What benchmarking/heuristics were used? Plotting the overall distribution of this metric and others used for filtering would be valuable.

All associations mentioned in the main text results should include the association statistics including p value and odds ratio. They should also all include their ClinGen classification score and category. Alternatively, these values could be provided in a main text table. For the HBB associations, statistics should be given for both the thalassemia and renal tract calcification phenotypes.

It is not clear why the authors dismiss the associations in UGGT2 and DNAH8. A rigorous statistical association should be sufficient even in the absence of functional data. There are many reasons why a true association in humans may not have experimental support in the literature.

The precise criteria used to define the seven 'compelling candidates' is not clear.

Supplementary table 2 should further stratify case and control numbers according to whether they are carrying an analysed variant in the stated gene (i.e. case carriers, case non-carriers, control carriers and control non-carriers).

According to Genomics England online documentation, the pipeline switched from GRCh37 to GRCh38 over 5 years ago, and only a minority of rare disease genomes are on GRCh37 so it's a bit surprising that the authors are still using a GRCh37 dataset.

Please can the authors include a description of the criteria used to define genes included in the original framework, which includes protein coding genes and RNA genes. Also was a single transcript tested or multiple transcripts, and how were those defined?

The authors should consider making full summary statistics available.

Minor comments

The introduction says that in silico triage highlighted 22 associations. Is this correct? This number does not appear to be mentioned elsewhere in the results.

How do the authors define CCR regions? There is a reference, but it would help the reader to have a sentence summarising it in this text.

On page 8 the authors define a high ClinGen score as >=8. This seems to conflict with the methods which defines moderate as 7-11 and strong as 12-18. Please can the authors clarify this.

A more detailed description or references are needed for the independent rare disease cohorts.

Regarding the COL4A3 association, please cite Solanki et al, 2023 that looks in detail at a variety of phenotypes associated with heterozygous COL4A3 variants.

I agree with the authors comment in the discussion that 'Further evidence is needed before all the novel associations described here can be used clinically for diagnostics, counselling and management'. Can the authors please comment in the discussion whether they feel that any of their associations pass this bar?

I find some of the statements such as 'confirmation would result in x new diagnoses' in the discussion and abstract to be too speculative. Those numbers could be limited to a higher confidence subset of associations.

The font size of figure 2 is too small to be readable. Perhaps this could be truncated to only include genes that pass a certain p value threshold. Alternatively, a Manhattan plot incorporating all phenotypes and genes might be clearer.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The revised version of the manuscript is clearer and more informative. The points I raised have been addressed or answered satisfactorily. I particularly appreciate the extended discussion.

However, I am surprised that the overlap between the 69 prioritised genes in the new analysis and the 88 prioritised in the first one is only 22. Also, only 2 of the 5 genes highlighted in the current version of the manuscript were included in the list of 88 from the first version of the manuscript. After readning the answers to reviewer 3 I understand that the current approach should be more appropriate. The fact that the ratio of previously known disease-gene associations compared to known+potential_new is higher might indicate it is indeed so, but I am not able to evaluate this from a statistics point of view.

Minor comment: the "G" in Cell A307 of tab "triage_burden_test_03_07_24" from Supp Table 3 I assume should be something else.

Referee #2

(Remarks to the Author)
The authors have addressed the majority of my comments. I will continue to defer to the other two reviewers on the updated statistical assessments. I have just a few minor remaining points that could be addressed.

With regard to GPR17, I might slightly reword the sentence “GnomAD evidence (o/e LoF = 0.79 (0.45-1.46) and pLI = 0) does not support a haplo-insufficient mechanism for this gene so the variants would have to be acting in a recessive or polygenic mechanism, the latter being more likely given the phenotype.” given that high LOEUF scores can also be seen in later onset, lower penetrance monogenic diseases as well. Perhaps “GnomAD evidence (o/e LoF = 0.79 (0.45-1.46) and pLI = 0) does not support a haplo-insufficient mechanism for this gene. This may be explained by the variants acting in a recessive manner, or the low LOEUF score may represent the later onset, incomplete penetrance aspects of the disease and/or a polygenic mechanism of disease, the latter being likely given the phenotype.”

For the sentence “However, further evidence is necessary before many of the novel associations described here can be used clinically for diagnostics, counselling, and management as a ClinGen classification of at least moderate evidence is required” I suggest adding “….for inclusion in diagnostic genetic test panels.” and reference Bean et al. Diagnostic gene sequencing panels: from design to report—a technical standard of the American College of Medical Genetics and Genomics (ACMG). 2020 Genet Med, 22(3):453-461. This publication was the basis of my prior comment.

I read Reviewer #1’s comment and the authors’ response: “L466 - it would be useful to have a description of the efforts done (and potential findings) through GeneMatcher. Have they considered other gene matching platforms or MatchMaker Exchange (which would give them access to larger numbers)?” and the response “The Genomics England research environment requires approval to orchestrate any potential findings with external researchers and GeneMatcher is the only provided option. This requires submission of a form per disease-gene association and manual review by their committee. So far we have performed this for all the highlighted associations but no significant collaborations have been identified through this to date. We agree it would be beneficial to have more widespread dissemination of our findings, leading to validation through new cases, and hope publication of this manuscript will support this goal.” Please note that GeneMatcher is connected to the federated Matchmaker Exchange (MME) such that entry of any gene candidate into GeneMatcher defaults to query of all other connected MME nodes as described in materials on our website https://www.matchmakerexchange.org/. However, I would also suggest that the list of candidates that have been classified and reached at least limited be submitted to GenCC. You can contact us at gencc@thegencc.org

For the new Figure 1, I would suggest including the breakdown of ClinGen gene classifications after “Application of ClinGen framework…” (e.g. 2 buttons with 42 limited and 27 moderate) and then have an arrow from the moderate bucket showing your criteria for highlighting the 5.

I had previously requested expansion of Supplemental Table 1 to include the full three lists of disease groups, subgroups and specific diseases and how they are related to each other ontologically which would allow other groups to apply the same categories if they wish. I appreciate that the authors have now provided the full list; however, the ontological mapping between the three lists has not been provided. In many cases, one could easily make assumptions about the relationships but in some cases, the mapping is less clear? Could the authors add one more column in this table and add the immediate

parent term, at least when there is only one parent term, which would allow the ontology to be replicated by others. I'm guessing some terms may map to more than one parent such as "undiagnosed monogenic disorder seen in a specialist genetics clinic (UMDSS)" in which case "multiple parent terms" could simply be noted.

Heidi Rehm

Referee #3

(Remarks to the Author)
Congratulations to the authors on the revised version of the manuscript. All my previous concerns have been satisfactorily addressed and I have no further comments.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Queen Mary University of London
William Harvey Research Institute
Barts & The London School of Medicine & Dentistry
Queen Mary University of London
Charterhouse Square
London, EC1M 6BQ

www.qmul.ac.uk/whri/
d.smedley@qmul.ac.uk

28th Sep 2024

Dear Nature Editorial Team and Reviewers,

We are very grateful for your constructive feedback and are delighted to submit a substantially revised version. Below **in bold** we detail our responses to all your reviews.

Yours faithfully,



Damian Smedley, Professor of Computational Genomics, QMUL

Editors comments:

Your manuscript, "Rare disease gene association discovery from burden analysis of the 100,000 Genomes Project data", has now been seen by 3 referees, whose comments are attached below. While they find your work of potential interest, as do we, they have raised important concerns that in our view need to be addressed before we can consider publication in Nature.

You will see that a common theme is the question of the advance over previous work - and in particular the Greene et al. study - and we ask that you address this concern comprehensively. Of course, also the comments about the robustness of the statistical tests must be addressed in full.

This revision process based on the valuable feedback from all reviewers has offered us a tremendous opportunity to improve our analytical approach. Briefly, we have changed the statistical test for the gene enrichment from the original one-sided Fisher's test to Firth's logistic regression model which is tailored to unbalanced case-control studies with rare events. This also allows correction for potential confounding factors to address the important comments about robustness. We have also added a more refined variant QC which is more tailored to rare Mendelian diseases and reviewed some testing criteria not to inflate unnecessarily the multiple testing as detailed below. As requested, our approach is now described more extensively in the methods and a detailed discussion of its pros and cons in the context of existing approaches as well as studies such as Greene et al has been added. We feel the valuable feedback from all reviewers, and more specifically the statistical comments and requests of reviewer 3, have allowed us to generate more confident associations that we now describe.

Should further experiments and/or analyses allow you to address these and all other criticisms, we would be happy to consider a revised manuscript (unless something similar has been accepted at Nature or appeared elsewhere in the meantime). We hope to receive your revised paper within four to six months. If you cannot complete the required revisions within this time frame, please let us know when you would anticipate being able to submit a revised manuscript.

Referees' comments:

Referee #1 (Remarks to the Author):

The manuscript presents the results from applying a gene burden analytical framework to discover rare disease-gene associations. Although the approach applied in the manuscript is not novel (see their reference 4), my understanding is that its extension, the development of an R package, its application to a large dataset of more than 35K cases and relatives, and the new associations presented are novel.

Comments:

L117-L122: consider if references 1 (Xiao et al) and 2 (Halley et al) are the most appropriate to indicate prevalence of RDs in the UK and the % of molecularly undiagnosed cases despite genome sequencing. For the latter, wouldn't it be more appropriate to directly refer to the diagnosis yield from 100K?

We agree and have now replaced reference 1 with a more accurate one that described prevalence worldwide and reference 2 with one referring directly to the 100K results.

L133-L141: unclear what is being presented in THIS manuscript for the first time (and is therefore novel) and what was previously developed and already published unless reading the referenced material (example: is the R framework completely novel or just its "extension"?)

We have modified this section to add wording about what is part of the prior study versus this study and to clarify the novelty here is the extension of the R framework to add refined rare variant filtering and a change of statistical modelling which are more tailored to Mendelian diseases and unbalanced case-control studies with rare events, making it generically applicable to any cohort, as well as new triaging and application to a larger 100KGP cohort to discover many new associations.

L155: authors claim that cases and controls could be defined in other ways, such as "by specific phenotypic annotations or phenotypic clustering". Have the authors tested this? If so, how different are the results compared with those from the approach they present in the manuscript? When is one approach more appropriate than another? How can a potential user decide?

This is also discussed in the comment below. In the 100,000 Genomes Project, strict recruitment criteria were developed to try and ensure the most similar patients were analysed using common virtual gene panels. Although considerable efforts were made to collect the most complete and precise phenotype annotations for each case, often this was not achieved which would limit a phenotype annotation or clustering approach. Finally, many different diseases have phenotypes in common so care needs to be taken with the phenotype-driven selection. We have explored both approaches for subsets of the 100,000 Genomes Project such as hearing loss patients and observed these pros and cons but it is hard to be conclusive on which approach is the best until novel findings are validated. We have now attempted to add a more nuanced discussion of these points: "The 100KGP is perhaps atypical compared to other projects in having a virtual panel-based approach for all types of rare diseases that required detailed inclusion and exclusion criteria for recruitment to specific disease categories allowing accurate case/control definitions. Phenotype-based selection strategies will be required when this is not the case and can overcome mis-recruitment but will require comprehensive and accurate phenotyping for maximum effectiveness."

L171, Fig 1 and Online Methods: it took me a while to figure out the "pipeline", since L171 and Fig1 start with 35,008 families, while Online methods start with 35,548. Might be useful to try to mention in the main text, or at least in Fig 1 legend, that this is a "filtered" cohort and refer to Online methods.

We have now added this clause to the sentence"..., filtered down from an initial total of 35,548 cases (see Methods), …"

L198-L198 (and other parts of the manuscript): consider discussing how some cases that might have previously received an incorrect or incomplete diagnosis (e.g. for a digenic situation) could be be limiting the power of the approach.

This is possible but unlikely and we have now explained this with the following addition: "It is possible but unlikely that criteria (iii) will have removed genuine associations due to incorrect/partial diagnoses as the Genomics England diagnostic pipeline is fairly conservative, generally requiring predicted LoF, *de novo* or known pathogenic variants in well-established genes for a diagnosis."

L205 - I am struggling to understand where the "87" comes from.

We apologise - this was a typo and should have been 88 but this has now been updated to 69 in the revised analysis.

L215-L218 - could these enrichments indicate these associations might be false positives?

This section on enrichment or not of particular ancestries is now removed as we have changed our statistical test to Firth's logistic regression with correction for many factors including inferred genetic ancestry.

L415-L416 - unclear; consider splitting in two sentences.

This whole section describing CKD and COL4A3 is now removed in the revised analysis.

L466 - it would be useful to have a description of the efforts done (and potential findings) through GeneMatcher. Have they considered other gene matching platforms or MatchMaker Exchange (which would give them access to larger numbers)?

The Genomics England research environment requires approval to orchestrate any potential findings with external researchers and GeneMatcher is the only provided option. This requires submission of a form per disease-gene association and manual review by their committee. So far we have performed this for all the highlighted associations but no significant collaborations have been identified through this to date. We agree it would be beneficial to have more widespread dissemination of our findings, leading to validation through new cases, and hope publication of this manuscript will support this goal.

L474 - given the assumption/limitation indicated here, would have it been more appropriate/better to define cases and controls "by specific phenotypic annotations or phenotypic clustering"? (see comment above).

As discussed above, although a strategy of defining cases and controls by specific phenotype annotations or clustering appears to be a better solution, the HPO annotations are often incomplete or lack precision. For example, for some recruitment categories the clinicians left out the primary phenotype and just entered phenotypes that would distinguish between subtypes of the disease. This actual sentence is now removed as the HBB example described is no longer significant in the revised analysis.

L603-L607 - unclear what is the rationale supporting the indicated limitations to "maintain statistical validity and power".

We agree these are somewhat heuristic thresholds and acknowledge they are arbitrary, yet this is commonly done in small dataset analyses. We have now edited it as : "To maintain statistical validity and power, the analysis was limited to those disease-gene associations where an arbitrary set of at least five cases exist…"

L639-L644 - seems gramatically incorrect. Also, a paranthesis seems to be missing (or there is one too many).

We thank the reviewer for spotting this error and have now corrected this sentence.

L646: "[...] or strong (12-18 points)" what? (missing a word here?)

Now corrected to clarify this refers to evidence of the association.

L646-L648: does not seem to be grammatically correct.

Now corrected to make clear why none of the associations can be classified as definitive as this point.

- Although the authors mention a study using BeviMed on approximately the same cohort, they have not compared their approach to other ones or discuss why theirs is better / more appropriate / Pros_Vs_Cons in comparison to other methods. A direct comparison with other methods would be useful but, if this is not possible or unjustified, some information for the reader to decide when/why to choose the authors approach would be appreciated.

We agree and have now described our analytical approach with more statistical details in the Methods. Briefly, the gene-based enrichment of variants in cases versus controls is quantified using the cohort allelic sums test (CAST) statistic under four proband's genotype scenarios. This is in turn assessed either within a one-sided Fisher's test as in our original submitted implementation or within a Firth's logistic regression model (current revised version) which is tailored to unbalanced case-control studies with rare events. The (Firth's) logistic regression model also allows correction for potential confounders. Importantly, we have expanded the discussion with advantages and limitations of our analytical approach in the context of existing approaches/software as well as studies such as Greene et al.:

"Rare variant gene burden approaches have been originally used in the context of complex genetics where a steadily increased number of (typically) unrelated individuals, now reaching

hundreds of thousands, are tested for genetic associations often across multiple quantitative and/or binary traits (i.e. genome-wide/pheno-wide association studies, GWASs/PheWASs). In this context, variants that confer *susceptibility* to a certain disease are of interest and, also depending on the applied allele frequency threshold for variant filtering, individuals contributing multiple rare variants per each gene are (typically) allowed in burden testing, this way increasing the chance to test variants with opposite effect (both *protective* and *deleterious*) within the same gene. Indeed, one of the main motivations behind the development of many of the existing gene burden tests beyond some basic implementations, e.g. the CAST statistic used within our *geneBurdenRD* approach, is the need to tackle the presence of variants with different effect directions and/or effect sizes within the same gene [37]. Selection of the input rare variants is crucial in any gene burden approach and can affect the statistical power, and becomes paramount in the context of Mendelian diseases, where we are not after multiple *susceptibility* variants (in any effect direction), but rather a single putative *disease-causative* variant (or pair of variants in a compound heterozygous genotype). Therefore, a scenario where an individual (whether case or control) contributes multiple rare genotypes to the same gene burden should not be contemplated. Our analytical framework (*geneBurdenRD*) is based on a convenient use of the established Exomiser variant prioritisation tool for rare Mendelian diseases to perform the non trivial step of variant annotation, scoring and segregation (using VCF files from single probands or family-based VCF files from nuclear families) and selection of the most predicted pathogenic, rare, segregating (putative *disease-causing*) variant(s) per each proband/family and each gene (and each compatible mode of inheritance). *geneBurdenRD* also provides case-control variant count filtering that mimics a Mendelian fully penetrant-like disease model with variants seen in controls not included in the analysis.

Among state-of-the-art software for whole-genome regression modelling of GWASs and PheWASs is REGENIE [38], that has been shown to be substantially faster than other similar approaches such as FastGWA [39], BOLT-LMM [40] and SAIGE [41]. In addition to several statistical tests which are mostly relevant to complex genetics, such as single-variant, gene-gene and gene-environment interaction tests as well as conditional analyses, REGENIE and some of these other related software can perform various gene burden tests, e.g., tailored tests to binary traits in both unbalanced and balanced case-control studies with rare variants, such as the saddle point approximation (SPA) approach [42] and the Firth's logistic regression model [43] (our statistical model of choice). Notably, both the SPA and the Firth's correction have been shown to provide good control of Type 1 error, but the SPA approach implemented in SAIGE resulted in inflated effect-size estimates, a feature which was not observed with the Firth's logistic regression in REGENIE [38]. Despite the ability to handle most commonly used input genetic and phenotypic files such as, for example, BGEN or PLINK/PLINK2 and tackle population structure and relatedness, REGENIE and similar approaches are not natively

tailored to the analysis of rare, putative *disease-causing*, *segregating* variants within families for rare Mendelian diseases. The crucial and often cumbersome step of annotation, scoring and selection of the pool of relevant rare variants to test for gene burden is left to the user as a preliminary step with the suggestion to use external tools such as snpEFF or VEP. Our analytical framework *geneBurdenRD* overcomes this limitation and covers an important need for more tailored gene burden analytical tools to the analysis of rare Mendelian diseases, with approaches like REGENIE remaining to be preferred for complex genetics analyses of single variants and beyond, across multiple quantitative and/or binary traits in (typically) unrelated individuals.

Although our approach applied to a large, rare disease cohort has successfully highlighted numerous known associations and suggested many previously unreported associations, it is not without limitations, pointing toward opportunities for future developments and re-analysis. For example, while we only considered single-nucleotide variants and small indels in protein-coding genes in the current study, the inclusion of non-coding rare variation as well as structural variants may help unveil further molecular diagnoses and disease mechanisms. Despite the large-scale size of the 100KGP, 29 of the 226 diseases observed in this analysis have less than 5 cases and a further 36 did not pass our testing criterium of at least two same disease cases with relevant variants in a certain gene. This highlights the current limited power in discovering novel (ultra-)rare monogenic disease associations and the need for even larger rare disease sequencing efforts. As the input to the statistical testing of each gene in our analytical framework is simply a matrix recording the presence or absence of a proband's genotype passing each criterion for each case and control, one intriguing possibility to increase statistical power is federated analysis across rare disease sequencing projects as such (preferably, Exomiser-based) processed data should be sharable under most ethical and legal frameworks. Finally, while our focus was rare Mendelian, monogenic fully-penetrant associations, further analytical developments and research are needed to uncover those missing genetic etiologies due to incomplete penetrance and/or variable expressivity as well as digenic and/or polygenic effects in rare diseases.

An alternative to our frequentist approach is offered by a Bayesian method for rare diseases called BeviMed [44] (Bayesian evaluation of variant involvement in Mendelian disease) in which disease risk depends on the genotypes at rare variants in a locus, a latent mode of inheritance and a latent partition of variants into pathogenic and non-pathogenic subsets. BeviMed was used in a recent analysis of approximately the same cohort of rare disease patients in this study where 241 known but only 19 novel associations [45] were found. 113/165 of our known and only 4/141 of the potentially novel (before triage) signals were also detected by BeviMed: *TUFT1* associated with epidermolysis bullosa; *SRP9* associated with ductal plate malformations, *MORC2* associated with mitochondrial disorders and *ARPC3*

associated with Charcot-Marie-tooth disease. Although a direct comparison is not straightforward due to the differences in the case-control selection strategies (the BeviMed study analysed 269 case sets as opposed to 161 that passed our testing criteria), let alone in the statistical approach, the relatively small overlap in signals highlights a possible complementarity of the two methods to discover new disease-gene associations from the same cohort."

- Can the framework be used (or easily adapted to be used) on federated datasets in which it is not necessary to have access to the full set of variants from each dataset? This would make it possible to increase the sample size without having to pool data centrally, overcoming many ethical, legal and technical challenges.

This is a very interesting suggestion that we had not properly considered. Our framework takes an input file of all the variants from the dataset and produces an intermediate file that summarises for each sample whether there is a relevant variant in each gene. It seems feasible that projects would be allowed to share such files under their ethical and legal frameworks and we could perform a federated analysis with a much larger N. We have now added to the discussion "As the input to the statistical testing of each gene in our analytical framework is simply a matrix recording the presence or absence of a proband's genotype passing each criterion for each case and control, one intriguing possibility to increase statistical power is federated analysis across rare disease sequencing projects as such (preferably, Exomiser-based) processed data should be sharable under most ethical and legal frameworks."

- It would be useful to define "proband" and "case".

Now added definitions at first mention.

- Fig 1: see comment above.

- Fig 2 unreadable in printed form. Consider splitting in more than one page or including only the blue (88 triaged novel signals) in this Figure and add the rest as supplementary.

We suspect the PDF conversion has reduced the resolution of the figures and hopefully we can work with the journal to ensure the final version is fine but in the new version the number of significant signals is also much fewer.

- Figs 3 and 4: unreadable in printed form. Try to improve readability and/or consider publishing only one or two of the main panels as a Figure and the rest as supplementary figures. Legends could be more detailed (ex. missing info about what is each subpanel and how to interpret it)

Again we suspect the PDF conversion has reduced the legibility of the figures and hopefully we can work with the journal to ensure the final version is fine. The new figures contain fewer panels and we have added more info on each subpanel to the legend as requested.

- Supp Fig 1: "highly" should be "high"? Also, abbreviations missing (BP, CC, MF).

This Supplementary Figure has now been removed as no longer relevant in the revised analysis.

The software developed and the outcomes of this manuscript are of interest to people in the fields of human genetics and bioinformatics.

The code of the open-source R framework (geneBurdenRD) and Exomiser are both available online. The installation and usage of geneBurdenRD seem to be well documented, although I have not tried to install and run it.

To the best of my knowledge, the manuscript does not have any flaws which prohibit its publication, the conclusions are original and the statistical tests seem appropriate, although I have indicated in my report where additional details might be useful for the reader.

We thank the reviewer for their feedback and feel the manuscript is now improved after responding to their comments.

Referee #2 (Remarks to the Author):

In this manuscript, the authors applied a gene burden analysis approach to a very large cohort of rare disease patients from the 100,000 Genomes Project and identified 88 novel gene-disease associations, in addition to identifying 190 previously known disease associations. They evaluate the evidence for each one according ClinGen's gene-disease validity framework, 38 of which meet moderate or strong levels of evidence with another 50 at limited. This demonstrates the utility of a large rare disease research datasets for delivering novel gene discoveries and candidates. Over all, this paper is a very nice body of work with major contributions to rare disease gene discovery. I have a few areas of feedback that, if addressed, could improve the paper.

We thank you Heidi for your positive comments and we feel we have now addressed your feedback and improved the paper.

There does not appear to be any list of the full set of variants that implicate the novel 88 gene-disease associations. This would be important to include so that the evidence is clear and readers and reviewers can better judge the full range of evidence for each novel gene-disease relationship and proposed mechanism of pathogenicity.

We have now included a Supplementary Table 5 that details the case variants driving the novel associations.

It's also not clear if each variant was separately classified according to pathogenicity versus only considering the total gene burden. I would recommend individual variant classification for those variants in the genes meeting strong or moderate ClinGen evidence. It would be useful to know how many individual variants reached pathogenic or likely pathogenic according to current ACMG/AMP standards versus how many were VUS on their own. I imagine some may have had sufficient evidence for reach P/LP based on predicted LOF impact, de novo occurrence, being in trans with another P/LP variants, etc. It's worth noting that ClinGen guidance suggests that once a gene-disease relationship meets strong or definitive, variants can be classified up to pathogenic, but if only moderate, the ceiling is LP and if limited, the ceiling is VUS. The variant classifications and supporting evidence should also be submitted to ClinVar so these variants are findable by the community.

We have also added the results of our automated Exomiser ACMG/AMP classification for all the variants to Supplementary Table 5, downgrading any P classifications to LP for moderate and LP to VUS for limited as recommended. It should be noted that our Exomiser

classification tends to under-classify pathogenicity as not all ACMG categories are easier to assess automatically. We are in the process of working with Genomics England to submit the P/LP variants to ClinVar.

The authors use the terminology “a ClinVar known pathogenic”. The authors have not defined what that description means and I would hope that it’s not defined by “one or more submitters to ClinVar called it pathogenic” as one cannot blindly trust all submissions to ClinVar as accurate. Instead, I would encourage the authors to classify variants according to the ACMG/AMP classification standards. Evidence and rationales submitted to ClinVar can be useful to contribute to variant classification, but an independent review of all evidence and entries in ClinVar is recommended.

This is a great suggestion and we have now switched to using our own ACMG classification (from the latest version of Exomiser) in the text (describing the ClinVar status as well where relevant).

The authors list “1 or 2 LOF” each for DYSF, KCNQ1, RYR1 and COL4A3 which have primarily missense variants, and in some cases, LOF is known to cause a distinct disease. Have the authors rigorously assessed the likelihood of rescue for these pLOF variants? A framework to guide this can be found in PMID: 37633279. Similarly, the pLOF variants found in gnomAD for UNC13A could be assessed with this framework.

Our Exomiser automated ACMG classification has implemented the recommendations in PMID: 37633279 when assessing PVS1 and in the revised text we now rely on this to establish whether pLOF variants are likely to escape NMD or not for each candidate.

The authors note that de novo annotations were restricted to “only trios or larger families where de novo calling was possible and provided by the user”. This seems like a missed opportunity to not do systematic de novo assessment on the call set for all trios. At a minimum, it would be useful to know how much of the data had de novo calling applied. I was particularly surprised to see that de novo variants were not observed in any of the novel associations. It would be important to note if this is due to only a small portion of the cohort having trios and de novo calling vs some distinction in the characteristics of the diseases for the novel gene-disease associations (age of onset, inheritance patterns, penetrance, etc).

We apologise for the confusion here and have now hopefully clarified this description. *De novo* calling was systematically done on all trios and larger families by Genomics England. 43% of cases were trios or larger families with *de novo* calls. A large proportion of these were

the intellectual disability cohort where there is a large amount of genetic heterogeneity and therefore it was unlikely that a new association explaining only 2 or 3 of these cases would reach significance. In the revised analysis 1/69 of the new signals are driven by *de novo* variants (CACNA1H and neurotransmitter disorders).

The authors noted "Potential associations were further filtered to 88/316 (28%) by removing those in which (i) the gene was a non-protein coding RNA one (31/316)". It would be useful to understand why non-protein coding RNAs were excluded. This seems like a missed opportunity for discovery.

Initially, this was a pragmatic decision to focus on protein-coding genes where there may be more reference knowledge and possibility of validation. We plan on further rounds of gene burden analysis where we will look at non-protein coding genes as well as non-coding and structural variants. We edited the manuscript as "Potential associations were further filtered and prioritised to".

For the 88 novel gene-disease associations, the authors note that they applied ClinGen's framework for gene-disease validity. This is excellent but I had a few questions/suggestions:
A) The section of the methods describing the process is labeled "Second clinical expert (GeCIP) and in silico triage" which does not really describe the gene-disease validity assessment. I would suggest labeling this section "Application of ClinGen framework for gene-disease validity" or something like that. Also, the first sentence notes that its automated, yet nothing described sounds automated! I will also note that its not clear how this step relates to the triage in general. Please clarify.

We agree the suggested title more accurately represents this section and have changed it. You are right that it was not all automated and have toned it down to "...was applied using *in silico* approaches where possible."

B) Was the evidence from the "5 interim confirmations" included in the gene-disease validity assessment or not? Just because there is a publication, does not mean the gene-disease association is now established so I would suggest you either include the published evidence in your primary assessment, or if not, then perform the assessment a second time with the published evidence included and make it clear what the current classification is. I might also refer to these published claims as "published gene-disease associations" or "published support" instead of "confirmations" which implies the mere fact of publishing something makes it definitive.

Absolutely - we realise a single other publication is not enough for an association to be confirmed and considered definitive and the efforts of ClinGen to resolve this are key to our field. We have now changed the wording throughout to "interim supporting evidence" instead of confirmations: the published evidence for the 5 interim associations was included in the ClinGen assessment, resulting in a score of 2 for the gene function fitting likely disease mechanism score.

C) It's also not clear what attributes placed 7 genes into the "compelling candidate" compared to other genes (subset of moderate??).

As also described below for reviewer 3, we have now reworded this section to try and be a bit clearer on the pragmatic rationale to select a reasonable number of candidates to "highlight" rather than really mean these are compelling and the others are not likely: "we chose candidates with prior functional data fitting the likely disease mechanism (Supplementary Table 4) and a ClinGen classification score (>= 8) to restrict to a manageable size of six candidates for further highlighting here."

D) On a related note, the authors state "Of the remaining 83 novel associations, nine had prior functional data fitting the likely disease mechanism and a high ClinGen classification score (>= 8)". Its not clear what "high" means here as this includes most of the spectrum of ClinGen's moderate evidence (7-11 points). I might suggest continued use of ClinGen's terms or using the score alone without the word "high" if you want to remain using 8 and above.

Corrected as in point (C)

Under the section of the paper describing COL4A3, the authors note that eleven of the cases already had PKD1 (or 2 with PKD2) diagnoses and raise the possibility of digenic inheritance or modifiers. After a quick look at the abstract of the referenced paper (citation 43) there may be some plausibility to this mechanism for COL4A3/COL4A4/COL4A5, given the heterotrimer configuration of these proteins. However, it's not clear if that is true for PKD1 and PKD2. Its also worth noting that PKD1 is one of the most polymorphic genes in the human genome with 1.8 million missense allele counts in the general population in gnomADv4 (average missense burden being 2.3 missense variants per person). Only 26 of the 5932 unique missense variants in gnomAD are classified as P/LP. As such, I would strongly suggest reviewing the detailed evidence on each variant before concluding digenic inheritance for any of these cases. None of the variants are shared for independent review.

In the new analysis based on a refined pool of rare variants and the Firth's logistic regression, adjusted for family size, age, inferred genetic ancestry and sex, this association is no longer significant at the stricter FDR threshold of 0.5% (the number of contributing cases dropped from 29 to 20; submitted 1-sided FRD-adjusted Fisher p-value = 7.5 x 10^{-6} vs revised covariate-adjusted, FDR-adjusted Firth's logistic p-value = 0.026) and the section has been removed.

The authors note "Further evidence is necessary before all the novel associations described here can be used clinically for diagnostics, counselling, and management." I initially interpreted this to say that none of the findings could be used diagnostically, etc., which I disagree with as some of the gene-disease associations meet strong or moderate, both levels being approved for inclusion in diagnostics panels. However, I'm now guessing the authors meant that not all of them are ready for use (e.g., the limited). Perhaps change "all" to "some" to make this clearer.

We thank you for pointing this out as we had not realised this was the current recommendation and have now changed the sentence to: "However, further evidence is necessary before many of the novel associations described here can be used clinically for diagnostics, counselling, and management as a ClinGen classification of at least moderate evidence is required."

In the Methods, the authors describe the selection of coding variants and removal of many other categories such as "coding transcript intron variant". Please clarify if intronic splice site variants (e.g., +/-1,2 and others in the splice consensus sequence) are included or excluded? If included, how far into the intron?

Canonical splice acceptor and donor, as well as splice region variants (either within 1-3 bases of the exon or 3-8 bases of the intron) are annotated separately and retained. This is now clarified in the methods.

The authors note "Each lolly indicates a variant by its protein change annotated on one single transcript (specified in the plot)…". Please confirm (and change the text to clarify) if these are MANE Select transcripts and if all variation is reported according to MANE Select (or MANE Plus Clinical if a stronger impact is predicted). If not, I'd recommend changing to this now established standard.

We now clarify in the methods that MANE Select (or MANE Plus Clinical if a stronger impact is predicted) is used.

Supplemental Table 1: In addition to providing the relative breakdown of cases across the disease groups, top subgroups and top specific diseases, it would be useful to simply provide the full three lists of disease groups, subgroups and specific diseases and how they are related to each other ontologically. I looked in the prior paper and could not find this articulated there either. This would allow other groups to apply the same categories if they wish.

In a new Supplementary Table 6 we now show these relationships and describe them in the methods.

Please use “variants” instead of “mutations” which is used occasionally in the paper.

We have now corrected the couple of mentions of mutations that had crept into the text.

Typos:
• KCNQ1 is misspelled as KNCQ1 at least twice.
• ClinVar is mis-capitalized as Clinvar in multiple places.
• In “Supplementary Table 1. Demographics of the 100,000 Genomics Project cohort.”, change “Genomics” to “Genomes”.
• In the sentence “Supplementary Figure 1. Gene ontology enrichment analysis of the GTEx substantia nigra gene co-expression network showing highly enrichment for terms associated with mitochondrial function, a key process in Parkinson’s disease pathophysiology.” “highly enrichment” needs a fix.

We have made all these typo corrections when still present in the revised text.

Heidi Rehm

Referee #3 (Remarks to the Author):

Key results and originality

In this work, the authors seek novel genes that contribute to rare diseases using gene burden analysis in the Genomics England 100KGP. They claim to have identified 88 novel associations, and highlight in detail a subset of seven that are particularly compelling. This is a very well written and clear manuscript. It addresses a critically important question, which is the need to identify more genetic causes of rare disease. It will be of interest to rare disease researchers and clinicians, and the seven highlighted associations are convincing.

We thank the reviewer for their kind words and confirming the utility to rare disease researchers worldwide.

The analytical framework described in this work uses exact tests to compare burden of pathogenic variants defined by Exomiser (Bone et al, 2016) in cases to controls. The framework was previously described (100,000 Genomes Pilot Investigators et al, 2021), and here the code has been adapted for generic application and published along with visualisation scripts.

Particularly thanks to the current revision, our gene burden analytical framework has been substantially refined with: (i) a more tailored rare variant filtering strategy that mimics Mendelian, fully penetrant-like disease models; (ii) a switch from one-sided Fisher's exact test to penalised likelihood-based Firth's logistic regression model which has been shown to provide bias-reduction for small sample datasets, particularly characterised by rare events, and allows adjustment for potential confounders; (iii) a reduction of the multiple testing as detailed below; (iv) the addition of useful visualisation scripts and (v) conversion to allow generic application to any large-scale cohort of rare disease samples. This is now released as a simple to use, open-source R-based framework that we hope will be of use in other projects with similar issues of large numbers of unsolved cases due to yet to be discovered disease-gene associations. The fact that the main input file for our analytical framework is simply a file that summarises for each sample whether there is a relevant variant in each gene opens up to the possibility of federated-like analyses with much larger N. We have now highlighted all these improvements and additions throughout the manuscript.

The 100KGP cohort itself has previously been published (100,000 Genomes Pilot Investigators et al, 2021). The previous publication included gene-based burden testing on ~57K genomes. The present study uses ~73K genomes, so I infer that there is a substantial overlap with those included in the previous publication.

This is correct.

A separate study (Greene et al, 2023) has been published on the same dataset and with a very similar aim (to identify novel disease – gene associations) albeit with a different statistical framework. The authors discuss this in the manuscript.

Given that previous publications have described Exomiser, the analytical framework, and attempts to identify novel disease – gene associations using the same (or a highly overlapping) dataset, in my view the novelty and impact of this study lies primarily in the newly claimed disease-gene associations.

As highlighted above and explained further below, the current revised framework is substantially updated and improved. However, we agree that apart from the improved framework that is now available as a generic, open-source R-based pipeline, and the 28% increase in the cohort size, the most important impacts are the newly described associations. In the NEJM, 2021 paper we only described "the most probable new, fully penetrant, mendelian" 22 associations and only 4 of these are in the new set of refined associations in this revision. This is due to changes in the statistical approach (mainly the refined pool of rare variants and the covariate-adjusted analysis) and a much more extensive, and in our opinion effective, triaging strategy with added clinical expertise via a highly collaborative effort with members of the Genomics England Clinical Interpretation Partnerships (GeCIPs).

Statistical robustness

I am concerned about the possibility that the novel disease-gene associations the authors claim to have identified could be enriched for false positives for the following reasons:
1) Surprisingly limited overlap with the findings of Greene et al. While as the authors point out, Greene et al use a different statistical framework and different case-control definitions, it is nevertheless a similar framework on the same dataset addressing essentially the same question, so I would expect a high degree of overlap in results.

We agree it would have been nicer to see more overlap but the fact that 61% (original submission) of our known associations were also observed by BeviMed does reassure us to some extent. Also, as shown below, even changing the statistical pipeline slightly results in many different significant signals so we are not too surprised by the levels of overlap with BeviMed on the 100KGP cohort. We suspect that small adjustments to the thresholds used by each tool would raise the overlap but we also hope that the two frameworks are different

enough to be complementary and maximise discovery. 7% of BeviMed's associations are novel compared to 32% of our final associations. It is possible one is underpowered or one is overly enriched for false positives but it is hard to be certain until these associations are validated or not. In the new analysis for this revision, 113/178 (63%) of our known associations were also observed by BeviMed along with 4 of the novel signals. As described for reviewer 1, we have now added a section comparing the pros and cons of our method relative to BeviMed and other gene burden approaches.

2) The authors identify 190 known and 316 novel (pre-triage) associations from their analysis. This ratio is surprising. From this kind of analysis, that includes a substantial proportion of diagnosed patients with pathogenic variants in known disease genes, I would expect the majority of significant associations to be known. This suggests that the novel associations may be enriched for false positives.

Regarding the proportion of diagnoses in this cohort, only 18% of the samples were fully/partially diagnosed at the time of the analysis and this has only risen to ~20% diagnosed in the latest data release of the 100KGP. So ~80% of the cases are going to be either: (i) not explained by monogenic genetic causes, (ii) explained by non-coding or structural variants in known disease genes or (iii) explained by a yet to be discovered disease-gene association. Unfortunately, it is hard to estimate the proportion of each of these, or the number of new disease-gene discoveries that we expect from a cohort of this size. However, we do agree that 316 novel vs 190 known from the initial statistical analysis may be surprising to some extent. As we discuss below, this is largely driven by the same variants being present in cases and controls in 179 out of 316 (57%) enriched signals which is unexpected unless there is incomplete penetrance or extensive phenotypic overlap between cases and controls. We have now performed analysis to confirm the latter is not the case (see point 4 below) and, importantly, we have added an extra variant QC to our analytical framework per each case-control analysis where "variants seen in at least 1 case (as heterozygote, homozygote or in a compound-heterozygote genotype) are discarded if seen in at least 1 control (as heterozygote/homozygote/compound-heterozygote, homozygote/compound-heterozygote, homozygote/compound-heterozygote, respectively) to mimic a Mendelian, fully penetrant-like disease model.", as detailed in Methods and Results.

3) The authors have not attempted to empirically validate the p value threshold they used to define significance.

As discussed below, we now present some data on recall and precision of known disease-gene associations under the different statistical tests and thresholds and discuss why we selected our final strategy.

4) 92% of known associations passed the authors filtering criteria (i) to (iii), but only 28% of novel associations did. This also suggests that the novel associations may be enriched for false positives.

This is indeed key and, as mentioned above, the main factor driving this difference was the 179/316 (57%) novel signals removed at triage due to > 33% of the case variants driving the signal appearing in the controls as well. This was the case in only 14/190 (7%) of the known associations. After this triage for case variants in controls we were comparing 137 novel and 176 known which is a lot more balanced. Inspired by the point raised by this reviewer and the observation that most removed signals at the first triage were due to variants in cases and controls, we now pre-filter the input pool of rare variants per each case-control analysis to remove all such variants as explained in point 2 above.

The authors could reassure readers on statistical robustness using the below approaches.

1) Investigating whether the p value distribution of their results is inflated by performing an n-of-1 permutation analysis to generate an appropriate null distribution of p values, and generating a QQ plot.

2) The authors could run an additional 'negative control' model using rare variants that are largely expected not to contribute to rare diseases. Eg predicted benign, or synonymous variants. Then compare the p value distribution to the models that use likely pathogenic variants.

3) With frameworks based on exact tests that cannot be corrected for covariates, care needs to be taken to harmonise case-control groups for factors such as age, sex, and ancestry to avoid spurious results. Please could the authors provide some reassurance that their case-control groups have been appropriately harmonised or corrected for relevant potentially confounding factors.

We thank the reviewer for their suggestions to ensure robust results. We believe the n-of-1 permutation has been used empirically in a context of complex genetics with ~39 billion statistical tests and ~19K binary and quantitative phenotypes where nearly all are expected to be under the null (Pubmed:34375979) which is quite a different scenario than our study in a context of a few hundred rare diseases, e.g. 1,855 of our 51,899 tests (~4%) are already known disease-gene associations and with a substantial proportion of new associations expected in the remainder. We have fully engaged with suggestions 2) and 3). As

described above, we have now opted for a penalised likelihood-based Firth's logistic regression (LR) model (https://www.jstor.org/stable/2336755; Pubmed:23788246; Pubmed:34017140), as implemented in the R package 'logistf' (see below), which replaces our original analysis using a one-sided Fisher's exact test. The Firth's approach not only allowed us to tackle the adjustment for potential confounders (age, sex, genetically inferred ancestry, and family size, as further mentioned in point 5 below), but has also been shown to provide bias-reduction for small sample datasets, particularly characterised by rare events, as is the case of this rare variant gene burden application, and to yield finite estimates even in case of separation (e.g. all individuals who carry a rare variant are cases). In the table below we provide a detailed description to guide the reviewer through all the statistical changes from the original submitted version to the final revised one, including recall and precision (for currently known associations using green genes in PanelApp only), and the estimate of the proportion of true null hypotheses (i.e. π_0) (obtained with R package qvalue and bootstrap method).

Remarkably, the covariate-adjusted Firth's LR model identified many more significant disease-gene associations compared to the (unadjusted) one-sided Fisher's test at the same submitted 5% FDR threshold (see table below, columns A and B). However, this is not surprising as, oftentimes, a covariate adjustment can bring cleaner signals. Importantly, this revision has offered us the opportunity to realise that we have been unnecessarily inflating the total number of statistical association tests by:

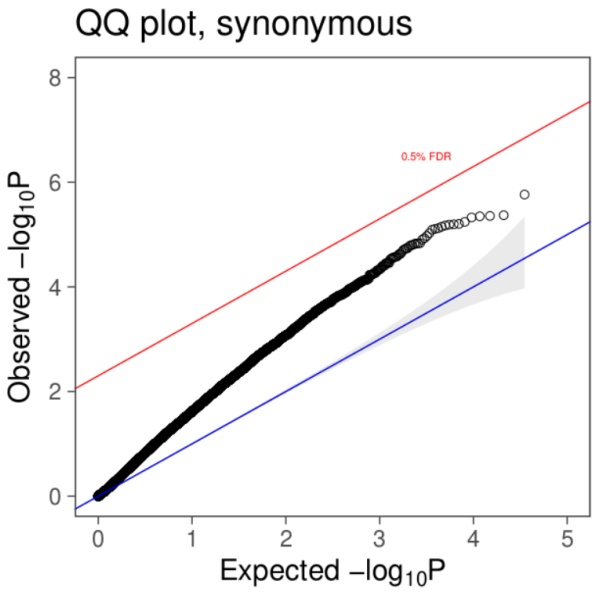
1. **including case segregating variants seen in controls (contributing to signals that we would discard right at the first triage as discussed above). With the additional variant QC described above that mimics a Mendelian, fully penetrant-like disease model, the total number of tests would reduce from 591,950 to 353,115 (see table below, column E);**
2. **including tests with only 1 contributing case (d == 1; 423,759 out of 591,950 - see table below);**
3. **including tests with gene-based enrichments more frequent in controls than in cases (putative *protective* as opposed to *disease-causing*) (63,323/591,950 - see table below).**

Points 2. and 3. were less apparent in the submitted version as those tests did not pass the significance threshold of 5% FDR with the one-sided Fisher's test (known to be more conservative). Since we would not prioritise signals driven by single cases at the first triage and we are not after *protective* genes in a Mendelian disease context (like for example in complex genetics), we have altered our testing criteria and adjusted the description in the paper as follows (new text is underlined): "at least five cases exist for the specific disease tested and, per each of the four gene-based proband's genotype scenarios above, where

relevant variants in the gene were seen in at least four probands, of which at least two were cases (we would not follow up associations signals driven by single cases/families in the first instance). Only gene-based enrichments that were more frequent in cases than controls (putative *disease-causing* as opposed to *protective*) were considered in the statistical correction for multiple testing.".

Altogether, applying the added Mendelian, fully penetrant-like variant filtering and testing only d > 1 and putative 'disease-causing' genes resulted in a substantial reduction of the number of the statistical tests (from 591,950 to 52,697, see table below, columns F and G). Finally, while triaging the new results, we found out that some of the triaged association signals were driven by the same variant in multiple cases due to cryptic relatedness. This led us to review further our sample QC and remove 157 related individuals (using kinship coefficients and a genetic variant overlap measure as documented in the Methods – "Pool of rare, putative disease-causing variants for gene burden testing") leaving a final QC-ed dataset of 34,851 samples and a total number of statistical tests of 51,899 (final revised version - see table below, column H).

As to the statistical significance threshold of the revised analysis, we used a pragmatic approach selecting a threshold of 0.5% for the FDR-adjusted p-values as it gave a reasonable recall of currently known associations (using green genes in PanelApp only) and a balanced number of known and novel signals (145 known out of 306) (see table below). In comparison with the submitted version (one-sided Fisher's test), this revised version (covariate-adjusted Firth's LR) retains similar recall (3.2% vs submitted 3.9%) and increased precision (47.4% vs 34.6%). Importantly, we tested for the enrichment of synonymous variants (using the same testing criteria of our new pipeline) as suggested by the reviewer and observed no signals at our chosen significance threshold (see QQ plot below; the red line indicates the 0.5% FDR threshold). A brief explanation of our threshold was added to the main results: "At this threshold, we observed a reasonable recall (3.2%) and precision (47.4%) of currently known associations (using green genes in PanelApp only), a balanced number of known and novel signals, and no enrichment at all for synonymous variants (empirical negative control burden test)".



We have corrected the known, novel and triaged signals throughout when using this new pipeline. After triaging, just over half of the novel signals corresponded to those described previously (including 2 of the 5 interim independent examples and 2 of the 7 "compelling" signals in the original submission). In addition, the new triaged associations contain 3 new interim independent examples and 3 new "compelling" examples that we now simply define as "highlighted" examples as explained below.

	A	B	C	D	E	F	G	H
	1-sided Fisher's test ***(original submitted version)***	Covariate-adjusted Firth's LR	Covariate-adjusted Firth's LR	1-sided Fisher's test	Covariate-adjusted Firth's LR with added variant QC	Covariate-adjusted Firth's LR with added variant QC, d>1, *protective* dropped	Covariate-adjusted Firth's LR with added variant QC, d>1, *protective* dropped	**Covariate-adjusted Firth's LR with added variant QC, d>1, *protective* dropped AND further sample QC with *'related'* dropped** ***(final revised version)***
***Significance* threshold**	**5% FDR**	5% FDR	1% FDR	1% FDR	1% FDR	1% FDR	0.5% FDR	**0.5% FDR**
Total tests	**591,950**	591,950	591,950	591,950	353,115	52,697	52,697	**51,899**
Tests with d == 1	**423,759**	423,759	423,759	423,759	281,685	n/a	n/a	**n/a**

Tests with *protective* associations	**63,323**	63,323	63,323	63,323	61,108	n/a	n/a	**n/a**
Total *significant* disease-gene associations	**506**	48,676	409	192	300	589	**353**	**306**
Known associations in PanelApp only (out of 4521 total tested)	**175**	831	163	126	126	193	**155**	**145**
Recall	**3.9%**	18.4%	3.6%	2.8%	2.8%	4.3%	3.4%	**3.2%**
Precision	**34.6%**	1.7%	39.9%	65.6%	42%	32.8%	43.9%	**47.4%**
0	**0.28**	0.30	0.30	0.28	0.39	0.47	0.47	**0.47**

4) How correlated are the studied diseases? Could the authors run a formal pairwise correlation analysis and presented it eg as a heat map, to a) reassure the reader that extensive phenotypic correlation isn't contributing to inflated p values and b) help identify treatment effects.

We have now performed this and include a new Supplementary Figure 1 and text in the results: "A pairwise phenotypic similarity comparison (Supplementary Figure 1) highlights that the highest levels of phenotypic similarity between samples are within these 'specific disease' categories, with extensive similarity often seen to other diseases within the same broad group (excluded in our control strategy), and no or more modest levels to the control disease categories (note that when combining all the control disease the median phenotypic similarity score between cases and controls was zero for all tested diseases). "

5) Plot the relationship between family size and number of analysed variants.

We now correct for potential differences in family size between cases and controls in our Firth's logistic regression approach (see above) as family size could indeed be a confounding factor for many case-control analyses. Indeed, in our previous NEJM paper the median number of filtered variants varies from 292 (258-327) for singletons to 149 (117-213) for duos and 29 (17-136) for trios and some diseases are enriched with singletons and duos rather than trios, signalling the importance of controlling the analysis for family size.

Additional data/information

Regarding ancestry:

1) Could the authors please replace self-reported ancestry with genomically-defined ancestry (eg using an algorithm such as Peddy). This would be more complete, accurate, and unbiased.

We have now replaced the self-reported ancestry with genetically-defined ancestry throughout the manuscript as requested. This was made available through analysis performed by Genomics England and we describe this in the methods.

2) Please include in the results section a description of the ancestral diversity of the cohort. This is currently in a supplementary table and should be more accessible.

We now include a description of the distribution of the genetically inferred ancestry in the results at the first mention of the cohort and Supplementary Table 1.

3) For the analysis of whether associations are linked to ancestry (page 7), please could full statistics (eg p value) be added.

This analysis and discussion are now removed as we adjust for the genetically inferred ancestry in our new pipeline.

4) Some of the associations have cases enriched for individuals of non-European ancestry. Can the authors please add some reassurance that those variants are enriched because of the disease and not because of the population structure?

Again, this section is now removed as we now adjust for the genetically inferred ancestry in our pipeline.

Please add a supplementary table detailing the 226 studied diseases including case and control numbers, and higher level categories.

We now include new Supplementary Tables 2 and 6 that contain this data. The relationship between the studied/specific diseases and higher level categories is now explicitly defined in Supplementary Table 6 as requested by reviewer 2.

Could the authors please explain how the threshold of 33.3% was selected, for the triaging filter (iii). What benchmarking/heuristics were used? Plotting the overall distribution of this metric and others used for filtering would be valuable.

As described above, we have now switched to a pipeline that removes all variants observed in both cases and controls before analysis to potentially decrease the number of false positives and mimic a Mendelian, fully penetrant-like disease model. Therefore, this threshold and filtering is no longer required in the triaging.

All associations mentioned in the main text results should include the association statistics including p value and odds ratio. They should also all include their ClinGen classification score and category. Alternatively, these values could be provided in a main text table. For the HBB associations, statistics should be given for both the thalassemia and renal tract calcification phenotypes.

The association statistics and ClinGen data are now added to each discussed association in the main text. The HBB signal is no longer significant under the Firth's LR model at 0.5% FDR.

It is not clear why the authors dismiss the associations in UGGT2 and DNAH8. A rigorous statistical association should be sufficient even in the absence of functional data. There are many reasons why a true association in humans may not have experimental support in the literature.

As discussed below, we have clearly confused readers with our description of "compelling" candidates and have now changed this to "highlighted" candidates throughout the text and improved our explanation of why we chose these to include as vignettes, i.e., the presence of existing functional supporting data. Indeed, we do not mean to dismiss the other associations or imply they are not strong candidates and we hope this is now clearer. In addition, the UGGT2 and DNAH8 associations are no longer significant in the new pipeline.

The precise criteria used to define the seven 'compelling candidates' is not clear.

On page 8 we describe how we chose candidates to further highlight based on those with the best evidence (ClinGen score >= 8) and prior functional data fitting the disease mechanism. We agree this latter categorisation is a bit unclear and have now extended the sentence to clarify this is indicated by bolding in the summary column of Supplementary Table 4: "Of the remaining 69 novel associations, we chose candidates with prior functional data for the gene fitting the likely disease mechanism (Supplementary Table 4) and a ClinGen classification score (>= 8) to restrict to a manageable size of six candidates for further highlighting here."

Supplementary table 2 should further stratify case and control numbers according to whether they are carrying an analysed variant in the stated gene (i.e. case carriers, case non-carriers, control carriers and control non-carriers).

This is now added.

According to Genomics England online documentation, the pipeline switched from GRCh37 to GRCh38 over 5 years ago, and only a minority of rare disease genomes are on GRCh37 so it's a bit surprising that the authors are still using a GRCh37 dataset.

We apologise for this error in our methods and have now corrected it to state that both GRCh37 (for the minority of earlier samples) and GRCh38 were used as part of the Genomics England alignment and calling pipeline that served as the input to our analysis.

Please can the authors include a description of the criteria used to define genes included in the original framework, which includes protein coding genes and RNA genes. Also was a single transcript tested or multiple transcripts, and how were those defined?

In the methods we have now clarified this with the following additions:

"The statistically significant associations were further filtered for those where (i) the gene was protein-coding as the Exomiser coding variant filtering settings also identified variants disrupting non-protein coding RNA genes (gene type definitions from the human gene nomenclature committee (HGNC) website),"

"The annotation of variants to genes comes from this Exomiser analysis using its default settings of identifying the most damaging consequence to the set of Gencode-basic tagged Ensembl transcripts."

The authors should consider making full summary statistics available.

We now include Supplementary Table 2 showing the full summary statistics for all tested associations (i.e. 51,899 tests over 161 case-control analyses), regardless of significance.

Minor comments

The introduction says that in silico triage highlighted 22 associations. Is this correct? This number does not appear to be mentioned elsewhere in the results.

We have now added text to clarify this is from the previous NEJM, 2021 study.

How do the authors define CCR regions? There is a reference, but it would help the reader to have a sentence summarising it in this text.

We have now added to the methods: "These CCR regions were defined previously by looking for the absence of variation in gnomAD[13] at various levels of certainty and here we used the 95% percentile download."

On page 8 the authors define a high ClinGen score as >=8. This seems to conflict with the methods which defines moderate as 7-11 and strong as 12-18. Please can the authors clarify this.

We understand how these differing thresholds can cause confusion. This was just a pragmatic way for us to select a reasonable number of candidates (<10) to highlight further in the main text. We hope we have now clarified this with the above changes to describing the strategy for selecting the six candidates.

A more detailed description or references are needed for the independent rare disease cohorts.

The lookup of variants in the independent rare disease cohorts mentioned in Supplementary Table 4 are described in the Methods.

Regarding the COL4A3 association, please cite Solanki et al, 2023 that looks in detail at a variety of phenotypes associated with heterozygous COL4A3 variants.

This example is now removed in the latest analysis.

I agree with the authors comment in the discussion that 'Further evidence is needed before all the novel associations described here can be used clinically for diagnostics, counselling and management'. Can the authors please comment in the discussion whether they feel that any of their associations pass this bar?

As discussed for reviewer 2, we have now clarified this: "However, further evidence is necessary before many of the novel associations described here can be used clinically for diagnostics, counselling, and management as a ClinGen classification of at least moderate evidence is required."

I find some of the statements such as ‘confirmation would result in x new diagnoses’ in the discussion and abstract to be too speculative. Those numbers could be limited to a higher confidence subset of associations.

We agree this is a stretch and have toned down the statement in the abstract to just be general “Further confirmation of these associations would lead to numerous diagnoses in ..” and in the discussion we have made it clearer this an upper bound of what would be possible.

The font size of figure 2 is too small to be readable. Perhaps this could be truncated to only include genes that pass a certain p value threshold. Alternatively, a Manhattan plot incorporating all phenotypes and genes might be clearer.

With the new analysis, the number of signals to show in figure 2 is much reduced and we hope this is now easier to read. The original figure version is quite sharp and some resolution may be getting lost in the PDF conversion as well.

Response to referees

Referee #1 (Remarks to the Author):

The revised version of the manuscript is clearer and more informative. The points I raised have been add ressed or answered satisfactorily. I particularly appreciate the extended discussion.

We are glad we addressed all your points and appreciate the feedback which has made this a much be tter paper.

However, I am surprised that the overlap between the 69 prioritised genes in the new analysis and the 8 8 prioritised in the first one is only 22. Also, only 2 of the 5 genes highlighted in the current version of the manuscript were included in the list of 88 from the first version of the manuscript. After readning the answers to reviewer 3 I understand that the current approach should be more appropriate. The fact that the ratio of previously known disease-gene associations compared to kn own+potential_new is higher might indicate it is indeed so, but I am not able to evaluate this from a stati stics point of view.

Yes - the pipeline is much improved after reviewer 3's suggestions in our opinion and we are even mor e confident of the new candidate associations we now present.

Minor comment: the "G" in Cell A307 of tab "triage_burden_test_03_07_24" from Supp Table 3 I assum e should be something else.

We have now corrected this.

Referee #2 (Remarks to the Author):

The authors have addressed the majority of my comments. I will continue to defer to the other two reviewers on the updated statistical assessments. I have just a few minor remaining points that could be addressed.

With regard to GPR17, I might slightly reword the sentence "GnomAD evidence (o/e LoF = 0.79 (0.45-1.46) and pLI = 0) does not support a haplo-insufficient mechanism for this gene so the variants would have to be acting in a recessive or polygenic mechanism, the latter being more likely given the phenotype." given that high LOEUF scores can also be seen in later onset, lower penetrance monogenic diseases as well. Perhaps "GnomAD evidence (o/e LoF = 0.79 (0.45-1.46) and pLI = 0) does not support a haplo-insufficient mechanism for this gene. This may be explained by the variants acting in a recessive manner, or the low LOEUF score may represent the later onset, incomplete penetrance aspects of the disease and/or a polygenic mechanism of disease, the latter being likely given the phenotype."

This is a great point and we have made this correction.

For the sentence "However, further evidence is necessary before many of the novel associations described here can be used clinically for diagnostics, counselling, and management as a ClinGen classification of at least moderate evidence is required" I suggest adding "….for inclusion in diagnostic genetic test panels." and reference Bean et al. Diagnostic gene sequencing panels: from design to report—a technical standard of the American College of Medical Genetics and Genomics (ACMG). 2020 Genet Med, 22(3):453-461. This publication was the basis of my prior comment.

Again, this makes perfect sense and we have made this correction.

I read Reviewer #1's comment and the authors' response: "L466 - it would be useful to have a description of the efforts done (and potential findings) through GeneMatcher. Have they considered other gene matching platforms or MatchMaker Exchange (which would give them access to larger numbers)?" and the response "The Genomics England research environment requires approval to orchestrate any potential findings with external researchers and GeneMatcher is the only provided option. This requires submission of a form per disease-gene association and manual review by their committee. So far we have performed this for all the highlighted associations but no significant collaborations have been identified through this to date. We agree it would be beneficial to have more widespread dissemination of our findings, leading to validation through new cases, and hope publication of this manuscript will support this goal." Please note that GeneMatcher is connected to the federated Matchmaker Exchange (MME) such that entry of any gene candidate into GeneMatcher defaults to query of all other connected MME nodes as described in materials on our website https://www.matchmakerexchange.org/. However, I would also suggest that the list of candidates that have been classified and reached at least limited be submitted to GenCC. You can contact us at gencc@thegencc.org

We have altered the text to make it clearer that the GeneMatcher submissions will allow full MME queries as well and have started the submission process to GenCC.

For the new Figure 1, I would suggest including the breakdown of ClinGen gene classifications after "Application of ClinGen framework…" (e.g. 2 buttons with 42 limited and 27 moderate) and then have an arrow from the moderate bucket showing your criteria for highlighting the 5.

We have enhanced Figure 1 as suggested.

I had previously requested expansion of Supplemental Table 1 to include the full three lists of disease groups, subgroups and specific diseases and how they are related to each other ontologically which would allow other groups to apply the same categories if they wish. I appreciate that the authors have now provided the full list; however, the ontological mapping between the three lists has not been provided. In many cases, one could easily make assumptions about the relationships but in some cases, the mapping is less clear? Could the authors add one more column in this table and add the immediate parent term, at least when there is only one parent term, which would allow the ontology to be replicated by others. I'm guessing some terms may map to more than one parent such as "undiagnosed monogenic disorder seen in a specialist genetics clinic (UMDSS)" in which case "multiple parent terms" could simply be noted.

We did include these mappings in a new Supplementary Table 6 described in the Online Methods which hopefully answers this question? The mapping is a simple hierarchy without multiple parents.

Heidi Rehm

Referee #3 (Remarks to the Author):

Congratulations to the authors on the revised version of the manuscript. All my previous concerns have been satisfactorily addressed and I have no further comments.

We are glad we managed to address your valuable feedback.

Yours sincerely,



Damian Smedley, Professor of Computational Genomics, QMUL