nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Gene-based burden tests of rare germline variants identify six cancer susceptibility genes
Corresponding author name(s): Dr Erna (Valdis) Ivarsdottir, Dr Kari Stefansson

Editorial Notes:

Transferred manuscripts This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Genetics.

Reviewer Comments & Decisions:

Decision Letter, initial version:

13th Sep 2023

Dear Dr Ivarsdottir,

Your Article, "Gene-based burden test yields rare germline variants in six novel cancer genes" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

We therefore invite you to revise your manuscript taking into account all reviewer comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Safia Danovi
Editor
Nature Genetics

Referee expertise:

Referee #1: cancer genetic epidemiology

Referee #2: cancer genetic epidemiology

Referee #3: cancer genetic epidemiology

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
Despite the success of GWAS of cancer much of the remaining heritability for most cancers remains to be elucidated. While larger GWAS meta-analysis are likely to harvest additional common risk variants, projects suggest that the possibility of gaining further traction is probably low. Even before the advent of GWAS there has always been the assertion that much of the inherited risk is likely to be enshrined in rare moderate penetrance alleles. With the availability of large sequencing projects testing this hypothesis is not possible. Here the authors have brought together three datasets identifying rare coding changes affecting the risk of several common cancers. The analysis is straightforward, and it is reassuring that breast cancer results accord with a recent larger study (Nat Genetics Easton and co-workers). The reduction in breast cancer associated with PPPIR15A is especially intriguing. The relation between ATG15 and colorectal cancer appears basically driven by a founder effect and its going to be interesting to see whether ongoing larger studies will also find support whether the relationship between coding changes in this gene and risk will really extend. My only real criticism of the paper is that it would have been nice if the data in Table 1 and thereafter would have been more accessible. I appreciate that its never as simple as GWAS cases and controls, and in this regard a web server as per the Astra Zeneca analysis of UK biobank would be helpful, certainly all of the data from Iceland needs to be readily accessible to the wider community.

Reviewer #2:
Remarks to the Author:
A. Summary of the key results

This paper has leveraged 3 large datasets and performed burden testing to identify novel genes related to cancer. The authors identified previous findings which reassures to a degree that the correct methodology was selected for this investigation. They claim 6 novel genes were discovered, with 2 genes having a protective effect. The association in PPP1R15A, suggests a protective effect against developing breast cancer with the authors claiming a 54% lower risk. The BIK finding is interesting, yet while the gene should not be considered novel in prostate cancer (authors quoted the findings from GWAS studies), the burden of rare variants is. The authors discover an interesting repeat that sits in the LCR region of BIK which is functionally predicted to cause various inframe deletions/insertions. The dosage effect of this finding (Supp Fig4) is another interesting aspect to this finding. The authors additional follow up one of the splice variants (p.S87G) discovered using the RNA sequencing data to confirm that this functions as a LoF. The authors also discovered an association for ATG12 in colorectal cancer. They also confirmed the LoF validity of a predicted start-lost variant. Additional, as this gene is in the vicinity of APC, the authors did not find any correlating variants

between ATG12 and APC, concluding independent associations from APC. Another interesting finding that the authors have discovered is in relation to CMTR2 and lung cancer. CMTR2 is a known lung cancer driver mutation (Intogen). While not mention by the authors, this could indicate a germline component to this known somatic driver. The authors performed a stratified analysis of various smoking characteristics which suggests that smoking status or intensity did not impact this association. They also report the association of CMTR2 within cutaneous melanoma. TG was reported for thyroid cancer and this burden in TG also appears to have nominally associations with serum levels of thyroid stimulating hormone. Lastly the authors report AURKB having a 16% lower risk with the diagnosis of cancer, which appeared to remain after removing high prevalent cancers.

B.Originality and significance
The originality here in my opinion is not large, essentially the authors have performed burden test on a much larger dataset. The originality here is that the sample size is very large. But, if it has taken over >480K WGS individuals to discover these associations, I am unsure how the findings of these genes will move the field forward, considering 2 of the genes are already known. I feel as the authors could have unpacked the findings in greater detail and had the opportunity to compare how these new findings fit in relation to previous findings (for example, what is the cumulative incidence in all variant carriers discovered per a disease outcome – not just the novel ones) and showed novelty in the approach to explain the significances of what these new findings tell us about each of the cancer sites and how this will move the field forward not to mention they have WGS, so other classes of variants could be examined.
The significance of this paper, while highlighting potential novel cancer related genes, I am left wondering how much more information these new findings explain in relations to the germline component of common cancers. It was unclear to me what the specific hypothesis of this paper was apart to dig deeper. The finding regarding PPP1R15A was, from my understanding, not performed in just females, hence this finding needs to be re-evaluated. If this does replicate, it may suggest in a subset of tumours that lose control of the ISR pathway inhibition may have some affect to PPP1R15A inhibition. May also present a preventively mechanism for future drug development. Yet other triangulated evidence would be required to make this claim. In the specific areas of research, the other associations are of interesting to those communities.

C.Data & methodology
Data:
The complexity of just the computational methodology to handle and process 497,083 WGS, not to mention the sequencing and the amount of expertise that would be involved in this project is phenomenal. This is no doubt why deCODE has the reputation of being the world's best genomics company. Especially in relation to germline genetics.
But I feel the authors have not provided adequate information regarding the specifics about the data and QC performed. In no particular order:
• As far as I am aware, the full release of the WGS data on the UK Biobank has not been released. There is no mention of what was done in this dataset apart from "Variant calling for all datasets was performed using GraphTyper" with a reference to the SV calling algorithm of Graphtyper. No paper or documentation was referenced on the UKBB in the genotype section. After spending time, I found that reference 98 needs to go into the genotyping section of the methods. Was the full cohort sequenced at deCODE, if not, documentation about how the other half of the cohort was created is needed. The authors need to mention how the data was harmonised. The overall burden of variants per a class between cohorts would be a recommendation.
• There is a lack of details around the sequencing QC steps performed specifically for this study with

the author quoting other papers (reference 98 needs to go here).
• Was the Norwegian study treated under the same conditions as the Icelandic population? Could you include a distributions plot of read depths between studies. Average coverage plots across chromosomes may be useful here as well.
• Excuse my inexperience here with Graphtyper - was maximum depth (down sampling) considered? Were any other metrics used apart from AD and DP for post-variant calling QC (was hard filtering/VQSR/random forests used?) –I feel another sentence in the "Variant selection for burden test and quality control" is needed for the justification of why just AD and DP was the only metric used. For heterozygous count, was there a method to determine variants didn't differ by DP/2, as I couldn't imagine you would get a perfect DP/2 for most variants. This section should come before the functional prediction. The sentence about the AAscore does not provide any information, a logistic regression model on what ?? just report that variants with a AAscore > 0.8 from GraphTyper was used.

• From my understanding, high-quality variants were selected using the mean counts to rule out somatic/artifacts calls. Since CMTR2 is a known driver – I suggest a plot showing that the variants calls are truly germline in nature. If somatic variation is observed for this gene (calls that fail the authors QC), this in my opinion is of great interest – It shows somatic change in the blood tissue outside the lung, also further supporting the authors findings in relations to little impact of smoking on this finding.
• The methods state 431,079 UKBB WGS were used, the introduction states 431,071 – which is why the number in the abstract is out by 8 compared to what the methods say.

Methods
Overall, my 2 biggest concerns here are related to the sex-specific effect of PPP1R15A and the lack of account for excessive relativeness/account for case-control imbalances. My recommendation would be to replicate the findings using SAIGE-GENE+ or Regenie.
• For the associations that were sex-specific cancer sites, why wasn't the burden test performed on just that one sex? If you see a ~50% reduction in developing breast cancer but 50% of the cohort is men (figure 1 – 19,121 vs 411,602 breast cancer for UK biobank, indicates men were included – male breast cancer is a different genetic disease) and the direct mechanism is acting via sex – I am unsure what this means in risk estimates and the biological context of this finding. Also why was the RNA-seq analysis not stratified by men only for the BIK finding?
• It was not clear to me how much of a role imputed variants played; a breakdown of this in the Supplementary Table 4 is needed. Also, a breakdown of this in the cancer data in ST1 (per cases and control status) would make it clearer. Is there a bias here because of the technology used? Would these associations go away without the use of imputed data? I think this information is critical in the interpretation of the results.
• As 55% of the variants that make up the BIK burden test come from a low copy repeat region, I would be extremely cautious in the interpretation of what this means in terms of a burden model. Could the authors make a comment regarding their justifications. Additionally, confirmation using long-read technologies (would be handy) to confirm the microsatellites are not a consequence of structural variation as there appears to be reported SVs in this region. While this finding is very interesting, more work needs to be done to unpack this.
• In the tables comment (line 450) *SAMHD1 was reported while this manuscript was being prepared. Also, the sentence in the text (line 119) that mentions this. This has been identified before this manuscript has undergone review. This should be removed as the paper is focused on novel findings.

• Excessive relativeness has not been addressed and the justification for this has not been given. It appears that first- and second-degree relatives were included to boost sample size. Yet the authors have not suggested that they have correctly adjusted for this either using a relationship matrix or other statistical frameworks. While the authors have used the LD intercept to adjust the X2 statistic, this is inadequate for rare variation. Please re-run the burden tests in Regenie/SAIGE-gene+ or other methods that can handle excessive relativeness.
• Additional QC or justification is needed on why low confidence pLoF sites were included. For example, on splice variants (CMTR2 – 16:71289350 both T and C alleles are reported as low confidence LoFs in gnomad)
• Why not necessary, it would be a nice touch for authors to justify why WGS is better to perform burden testing for LoF/missense compared to WES, which in theory should be able to identify these classes of variants. As this is the largest WGS burden study to date.
• In the stratified analysis of various smoking characteristics, could the authors provide a breakdown of carrier status for CMTR2 LoF/missense.

Quality of presentation

Overall, the structure of the paper is fine, some parts in the methods need readjusting (mention above). The presentation of the figures is very clear. Figure 1 is presented very well. In my opinion, I think supplementary figure 3 should go in the main figures. Thank you for including QQ-plots.

Appropriate use of statistics and treatment of uncertainties
This was already mentioned in the methods.

D. Conclusions:
My overall conclusion is that this paper is suited for another journal after addressing the concerns raised. The findings while interesting to a degree, I feel the authors could have done a lot more with the resource they had in relation to WGS and sample size. The robustness and validity of these findings are questionable due various issues that has been mentioned before. The reliability is reasonable, with the findings to a degree replicating in Iceland data which suggest the same areas are impacted by different variation which would make sense.

F. Suggested improvements: experiments, data for possible revision
Major:
• Perform the burden test in a software package that handles excessive relativeness.
• Perform burden analysis in women only for breast cancer.

Minor:
• More details surrounding sequencing/QC/LoF filtering.
• Is CMTR2 variants related to the somatic CMTR2 variant?
• Long-range sequencing/nanopore on the BIK finding
• Explanation of how much more information these new findings add? Do they explain more to the missing heritability?
• How these findings relate to other large burden studies such as genebass.org and how the use of WGS for exonic variation is justified.

• Burden analysis of germline variants that influence non-coding elements (personally would love to see a burden of CpG sites or mQTL sites). You should see BIK again for prostate.

G. References: appropriate credit to previous work?
Apart from reference 91, I feel the references are appropriate.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Overall, I think the lucidity is ok.

I think the clarity/context could be improved by these minor comments:
• Abstract – "is one way to shed light" is too colloquial in my opinion.
• 2 of these genes are not novel.
• "Protection of cancer" – associate with a decreased risk of cancer?
• The number of WGS is different to what is stated in methods.
• Introduction, while the introduction is short, I feel the size is appropriate.
• The hypothesis of the study is a bit unclear.
• A sentence that follows up from the founder populations statement into the specifics of the study design I think is needed.

Reviewer #3:
Remarks to the Author:
This paper presents the results of an exome-wide association analysis for all the major cancer types, using data from the datasets (UK Biobank and datasets from Iceland and Norway). The analyses identify six "novel" associations using burden analyses based on LoF and LoF+deleterious missense variants, two of which are protective (one for all cancer).

This is a very valuable analyses, such analyses have not been done for all cancers previously. While it is striking that most of associations were already known (something worthy of comment in the discussion), the identification the novel associations are interesting biologically and some could have therapeutic relevance.

The analyses follow a standard approach and are appropriate. One could argue that the Bonferoni correction is not sufficiently stringent since more than 20 cancers endpoints were considered, but this is always a somewhat arbitrary decision.

My main comment on the analysis relates to the absolute risk estimation. It is fairly meaningless derive these estimates within the cohorts since, as is pointed out, the recruitment e.g., for UK Biobank over a limited age range. If absolute risks are to be presented, it would be better to apply the estimated ORs to national incidence rates.

As per Nature Genetics guidelines, it is important that the full genome-wide set of summary statistics is deposited in a public repository, e.g., the GWAS Catalog.

Other specific comments

1. Introduction. It's not clear that the founder effect argument is relevant here, given that the largest component of the data is from UKB.
2. The definition of likely deleterious missense variants seems reasonable. However, some justification as to why these particular tools were used – there are many other possibilities of course.
3. The SAMHD1 association has been published (or at least is in the public domain) – so it should be included among the previously described genes.
4. PPR1R15A– (l125) – over two-fold protection is an exaggeration - approximately two-fold reduction in risk would be more accurate.
5. Related to this, while the associations are quite robust, the risk estimates are much more uncertain, in particular due to the potential for winner's curse. This should be acknowledged in the discussion.
6. It appears that both incident and prevalent cases were included in the analyses, please clarify in the methods. Were only cancers identified through linkage to cancer registries included or were self-reported cancers also included?
7. BIK (l164) what does "driven" mean here? What happens if p.R618H is excluded. More generally, is there evidence for any of the reported genes that the effect size varies by variant?
8. Phenome-wide associations (l311). This does not seem quite the right analysis, because the power for different cancers varies hugely. I think you need to do a test of heterogeneity of the OR among cancer types, and/or evaluate the association for all other cancers combined, after the identified associations are removed.

Author Rebuttal to Initial comments

Reviewer #1:

Remarks to the Author:

Despite the success of GWAS of cancer much of the remaining heritability for most cancers remains to be elucidated. While larger GWAS meta-analysis are likely to harvest additional common risk variants, projects suggest that the possibility of gaining further traction is probably low. Even before the advent of GWAS there has always been the assertion that much of the inherited risk is likely to be enshrined in rare moderate penetrance alleles. With the availability of large sequencing projects testing this hypothesis is not possible. Here the authors have brought together three datasets identifying rare coding changes affecting the risk of several common cancers. The analysis is straightforward, and it is reassuring that breast cancer results accord with a recent larger study (Nat Genetics Easton and co-workers). The reduction in breast cancer associated

with PPPIR15A is especially intriguing. The relation between ATG15 and colorectal cancer appears basically driven by a founder effect and its going to be interesting to see whether ongoing larger studies will also find support whether the relationship between coding changes in this gene and risk will really extend. My only real criticism of the paper is that it would have been nice if the data in Table 1 and thereafter would have been more accessible. I appreciate that its never as simple as GWAS cases and controls, and in this regard a web server as per the Astra Zeneca analysis of UK biobank would be helpful, certainly all of the data from Iceland needs to be readily accessible to the wider community.

Response 1:

We feel that we have done our best to make the results accessible to the reader: i) by showing the combined overall burden results for novel risk genes in Table 1, ii) by showing result for individual key variants per novel risk gene in Table 2, and iii) by summarizing burden association analysis effect estimates for all significant genes in our study in Figure 2 and Supplementary Figure 2. We realize that we are analyzing and presenting results for large datasets, however we do not have the resources to design and maintain an online tool to present our results.

The genotypes of the 500K WGS individuals from the UK Biobank has now been released and the summary statistics of our meta-analysis will be made available at https://www.decode.com/summarydata/.

Reviewer #2:

Remarks to the Author:
A. Summary of the key results

This paper has leveraged 3 large datasets and performed burden testing to identify novel genes related to cancer. The authors identified previous findings which reassures to a degree that the correct methodology was selected for this investigation. They claim 6 novel genes were discovered, with 2 genes having a protective effect. The association in PPP1R15A, suggests a protective effect

against developing breast cancer with the authors claiming a 54% lower risk. The BIK finding is interesting, yet while the gene should not be considered novel in prostate cancer (authors quoted the findings from GWAS studies), the burden of rare variants is. The authors discover an interesting repeat that sits in the LCR region of BIK which is functionally predicted to cause various inframe deletions/insertions. The dosage effect of this finding (Supp Fig4) is another interesting aspect to this finding. The authors additional follow up one of the splice variants (p.S87G) discovered using the RNA sequencing data to confirm that this functions as a LoF. The authors also discovered an association for ATG12 in colorectal cancer. They also confirmed the LoF validity of a predicted start-lost variant. Additional, as this gene is in the vicinity of APC, the authors did not find any correlating variants between ATG12 and APC, concluding independent associations from APC. Another interesting finding that the authors have discovered is in relation to CMTR2 and lung cancer. CMTR2 is a known lung cancer driver mutation (Intogen). While not mention by the authors, this could indicate a germline component to this known somatic driver. The authors performed a stratified analysis of various smoking characteristics which suggests that smoking status or intensity did not impact this association. They also report the association of CMTR2 within cutaneous melanoma. TG was reported for thyroid cancer and this burden in TG also appears to have nominally associations with serum levels of thyroid stimulating hormone. Lastly the authors report AURKB having a 16% lower risk with the diagnosis of cancer, which appeared to remain after removing high prevalent cancers.

B.Originality and significance

Q 2.1 **The originality here in my opinion is not large, essentially the authors have performed burden test on a much larger dataset. The originality here is that the sample size is very large. But, if it has taken over >480K WGS individuals to discover these associations, I am unsure how the findings of these genes will move the field forward, considering 2 of the genes are already known. I feel as the authors could have unpacked the findings in greater detail and had the opportunity to compare how these new findings fit in relation to previous findings (for example, what is the cumulative incidence in all variant carriers discovered per a disease outcome – not just the novel ones) and showed novelty in the approach to explain the significances of what these new findings tell us about**

each of the cancer sites and how this will move the field forward not to mention they have WGS, so other classes of variants could be examined.

Response 2.1:

Regarding "**BIK should not be considered novel in prostate cancer".**

We respectfully disagree with the reviewer regarding the novelty of *BIK* being a prostate cancer susceptibility gene. Even though other variants, located near *BIK*, have been reported to be associated with prostate cancer, no proof of a direct effect of these variants on the gene function have been provided, to the best of our knowledge. However, the burden of rare coding variants in *BIK* associating with prostate cancer and the effect on expression levels shown by us in this manuscript are the first direct implication of *BIK* as a prostate cancer susceptibility gene.

Regarding "**2 of the genes are already known**".

We are not sure what other gene the reviewer is referring to. However, in the paper, we describe genes as novel if we find association between burden of rare coding variants in the gene and cancer, if rare coding variant association in the genes have not been previously reported. We have now clarified our interpretation of "novel" in line 137 in the main text:

"For six of the 34 genes, no rare coding variants had previously been reported to associate with cancer risk and we describe them as novel cancer genes"

Regarding: "**But, if it has taken over >480K WGS individuals to discover these associations, I am unsure how the findings of these genes will move the field forward.**"

Detecting rare disease variants has low statistical power and one way to increase power is to perform analyses on very large sample sizes. Rare protective variants in particular require very large sample sizes to be found. In our opinion, the discovery of associations between loss-of-function variants in novel genes and cancer risk can advance the field in several different ways: i) the results can be utilized to identify high-risk individuals and thereby impact cancer screening and cancer prevention strategies; ii) the results can provide novel biological insight into the pathogenesis of cancer and

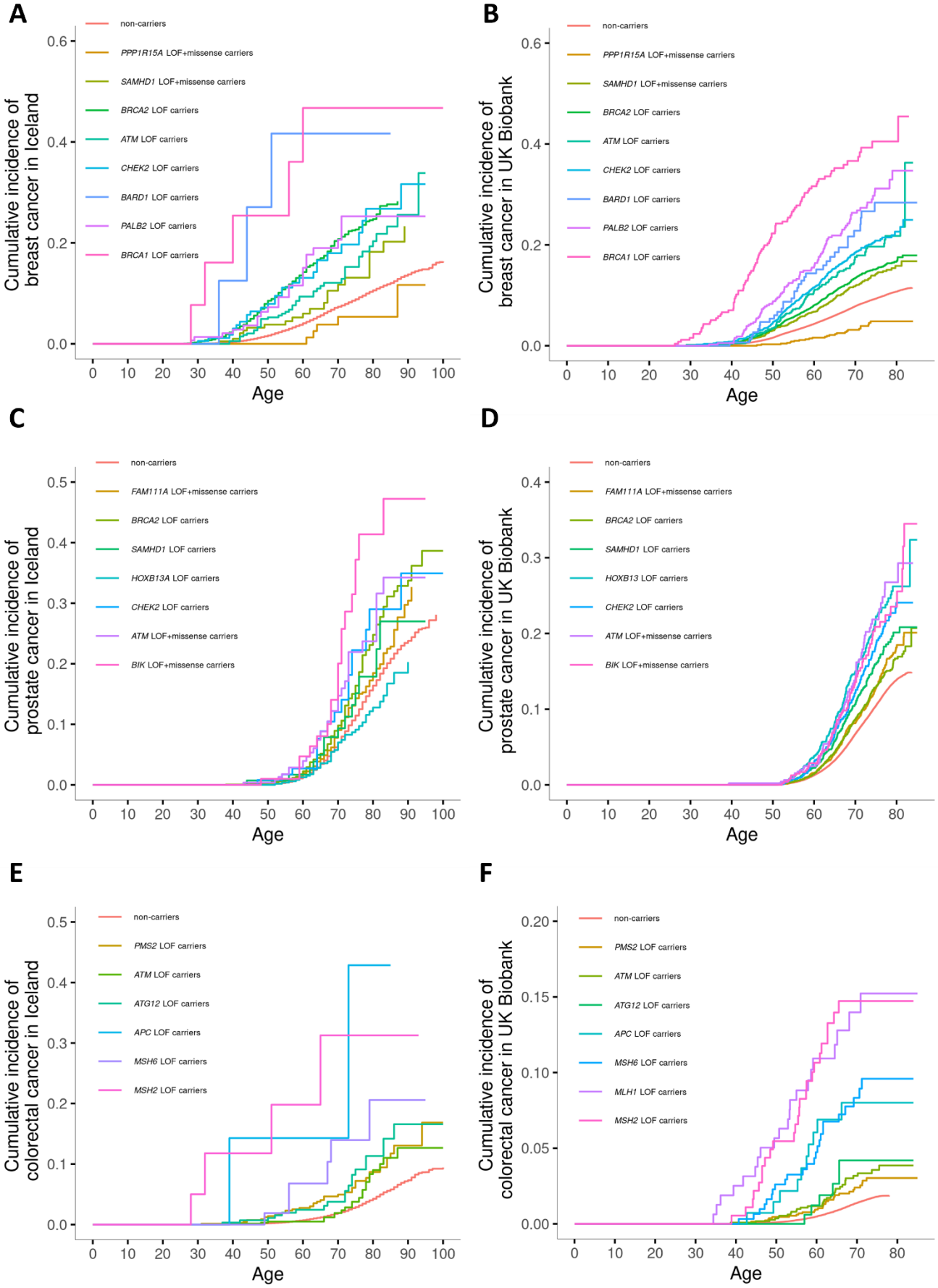
thereby point to possible new drug targets, especially variants that demonstrate that the loss of the function of a particular gene results in substantial protection against a cancer. While most clinical drug candidates fail due to efficacy issues, validation through human genetics is a strong indicator of the success of drug development[1] and association between loss-of-function variants and diseases has successfully predicted the safety and effects of pharmacological inhibition for several genes[2].

Regarding: "**I feel as the authors could have unpacked the findings in greater detail and had the opportunity to compare how these new findings fit in relation to previous findings (for example, what is the cumulative incidence in all variant carriers discovered per a disease outcome – not just the novel ones) and showed novelty in the approach to explain the significances of what these new findings tell us about each of the cancer sites**"

We agree with the reviewer that it would have been tempting to dive into each of these new findings in greater detail. However, we are reporting several new cancer risk loci for multiple cancer sites and are already struggling to fit our manuscript into the format/length as instructed by the Journal. Furthermore, we fear that we will have to shorten it even more before publication and are therefore only willing to extend its length dramatically per editorial instructions.

In Figure 2, we show how the effects of these new findings fit in relation to previous findings. The figure shows that *PPP1R15A* and *AURKB* are the only protective associations found, that *CMTR2* is the only gene associated with lung cancer, and that the other novel risk increasing genes, all have high effects comparable to previously known genes.

As the reviewer suggested, we have now also added Supplementary Figure 4 (shown below) demonstrating the cumulative incidence curves for all genes that were associated with breast cancer, prostate cancer and colorectal cancer.


A
non-carriers
PPP1R15A LOF+missense carriers
SAMHD1 LOF+missense carriers
BRCA2 LOF carriers
ATM LOF carriers
CHEK2 LOF carriers
BARD1 LOF carriers
PALB2 LOF carriers
BRCA1 LOF carriers
Cumulative incidence of breast cancer in Iceland
Age
B
non-carriers
PPP1R15A LOF+missense carriers
SAMHD1 LOF+missense carriers
BRCA2 LOF carriers
ATM LOF carriers
CHEK2 LOF carriers
BARD1 LOF carriers
PALB2 LOF carriers
BRCA1 LOF carriers
Cumulative incidence of breast cancer in UK Biobank
Age
C
non-carriers
FAM111A LOF+missense carriers
BRCA2 LOF carriers
SAMHD1 LOF carriers
HOXB13A LOF carriers
CHEK2 LOF carriers
ATM LOF+missense carriers
BIK LOF+missense carriers
Cumulative incidence of prostate cancer in Iceland
Age
D
non-carriers
FAM111A LOF+missense carriers
BRCA2 LOF carriers
SAMHD1 LOF carriers
HOXB13 LOF carriers
CHEK2 LOF carriers
ATM LOF+missense carriers
BIK LOF+missense carriers
Cumulative incidence of prostate cancer in UK Biobank
Age
E
non-carriers
PMS2 LOF carriers
ATM LOF carriers
ATG12 LOF carriers
APC LOF carriers
MSH6 LOF carriers
MSH2 LOF carriers
Cumulative incidence of colorectal cancer in Iceland
Age
F
non-carriers
PMS2 LOF carriers
ATM LOF carriers
ATG12 LOF carriers
APC LOF carriers
MSH6 LOF carriers
MLH1 LOF carriers
MSH2 LOF carriers
Cumulative incidence of colorectal cancer in UK Biobank
Age

Regarding: “**so other classes of variants could be examined.**”

We agree with the reviewer that exploring other classes of variants would be interesting, but in this current study we have decided to focus our analysis on loss-of-function- and predicted deleterious coding variants.

Q 2.2 **The significance of this paper, while highlighting potential novel cancer related genes, I am left wondering how much more information these new findings explain in relations to the germline component of common cancers. It was unclear to me what the specific hypothesis of this paper was apart to dig deeper. The finding regarding PPP1R15A was, from my understanding, not performed in just females, hence this finding needs to be re-evaluated. If this does replicate, it may suggest in a subset of tumours that lose control of the ISR pathway inhibition may have some affect to PPP1R15A inhibition. May also present a preventively mechanism for future drug development. Yet other triangulated evidence would be required to make this claim. In the specific areas of research, the other associations are of interesting to those communities.**

Response 2.2:

Regarding “**I am left wondering how much more information these new findings explain in relations to the germline component of common cancers**”.

See Response 2.1 above.

Regarding “**It was unclear to me what the specific hypothesis of this paper was apart to dig deeper**”.

Our hypothesis can be referred to as classical within the field of cancer genetics, i.e. that mutations in coding regions of genes confer risk of cancer. By leveraging large WGS datasets and performing burden tests, we enhance our statistical power, allowing us to discover several novel rare coding variants that confer risk of cancer.

We now aim to make the hypothesis clearer in the introduction section of the main text (line 103) by saying:

“To search for rare germline coding variants associating with cancer risk, we utilized three large genetic datasets…”

Regarding *PPP1R15A*, the association between burden of LOF+missense variants in *PPP1R15A* does not change when we restrict to using only females. In particular, the OR is 0.47 (95% CI = [0.35-0.63]) using both female and male cases, and using only females the OR is 0.48 (95% CI= [0.36-0.65]). For more details see Response 2.10 below.

C.Data & methodology

Data:

Q 2.3 **The complexity of just the computational methodology to handle and process 497,083 WGS, not to mention the sequencing and the amount of expertise that would be involved in this project is phenomenal. This is no doubt why deCODE has the reputation of being the world’s best genomics company. Especially in relation to germline genetics.**

But I feel the authors have not provided adequate information regarding the specifics about the data and QC performed.

Response 2.3:

We thank the reviewer for his/her kind words towards deCODE and agree with the reviewer’s comment about the fact that important information was missing from Methods section. We have now changed the Genotyping subsection in Methods, by including more detailed information about the genotype data and the QC performed:

“The genome of the Icelandic population was characterized by whole-genome sequencing (WGS) 63,460 Icelanders using Illumina technology, including GAIIx, HiSeq, HiSeqX, and NovaSeq machines, with subsequent long-range phasing[91] and imputation into 173,025 chip-typed individuals employing Illumina SNP chips, as has been previously described[13,92]. The 63,460 individuals were sequenced with a total of 192,637 lanes, 175,016 (90.85%) of which were prepared without PCR (Supplementary Table 16). Duplicated sampled were discarded based on sequencing yield and only samples with a genome-wide average coverage of 20x and higher were included. We

used read_haps[93] to detect sample contamination and removed contaminated samples. The average genome-wide sequencing coverage was 39.8x (sd 14.2, min:20.0x, max:397.8x). Joint variant calling was performed using GraphTyper (v.2.7.1)[94]. Further, familial imputation of genotypes in first- and second-degree relatives was used to increase sample size[13].

The Norwegian genotype data were generated at deCODE by WGS 2,576 individuals of Norwegian origin and subsequent imputation into 45,768 chip-typed Norwegians belonging to various research projects being worked on in collaboration with deCODE genetics (described above). Long-range phasing and sequence variant calling were performed jointly for multiple cohorts of various other origin (Danish, N-American, Iranian, Swedish and Dutch) where 50,839 have been WGS and 1,041,174 chip-typed. The 2,576 Norwegians were sequenced using Illumina HiSeqX and NovaSeq sequencing machines with a total of 8,034 lanes and all samples were prepared without PCR (Supplementary Table 16). The WGS protocol for the Norwegians was the same as described above for the Icelanders. The average genome-wide sequencing coverage was 37.2x (sd 6.3, min:20.1x, max:98.5x) and joint variant calling was performed using GraphTyper (v.2.7.5)[94].

The UK Biobank genotype data consists of 431,079 WGS white British/Irish individuals (ethnicity was determined by principal component analysis). The WGS was performed using Illumina NovaSeq sequencing machines at deCODE for 214,548 individuals and Wellcome Trust Sanger Institute for 216,531 individuals. The protocol for a preliminary release of the WGS of the UK Biobank dataset has been previously described in detail[95] and the additional samples analyzed here were prepared and sequenced in the same way. The average genome-wide sequencing coverage was 32.4x (sd 4.1, min:22.1x, max:162.1x) and joint variant calling was performed using GraphTyper (v.2.7.5)[94].

The Danish genotype data were generated at deCODE by WGS 1,932 individuals and subsequent imputation into 377,448 chip-typed individuals from CHB and DBDS (described above). Long-range phasing and sequence variant calling were performed jointly for multiple cohorts of various other origin (Norwegian, N-American, Iranian, Swedish and Dutch) where 50,839 have been WGS and 1,041,174 chip-typed. The WGS protocol for the Danes was the same as described above for the Icelanders and joint variant calling was performed using GraphTyper (v.2.7.5)[94]."

In no particular order:

Q 2.4 **• As far as I am aware, the full release of the WGS data on the UK Biobank has not been released. There is no mention of what was done in this dataset apart**

from "Variant calling for all datasets was performed using GraphTyper" with a reference to the SV calling algorithm of Graphtyper. No paper or documentation was referenced on the UKBB in the genotype section. After spending time, I found that reference 98 needs to go into the genotyping section of the methods. Was the full cohort sequenced at deCODE, if not, documentation about how the other half of the cohort was created is needed. The authors need to mention how the data was harmonised. The overall burden of variants per a class between cohorts would be a recommendation.

Response 2.4:

The full WGS data for UK Biobank has now been released. The protocol for a preliminary release of the WGS of the UK Biobank dataset has been described in detail and the additional samples analyzed here were prepared, sequenced and harmonized in the same way as the preliminary release. We have now added this information to the Methods section as described in Response 2.3 above.

Q 2.5 **• There is a lack of details around the sequencing QC steps performed specifically for this study with the author quoting other papers (reference 98 needs to go here).**

Response 2.5:

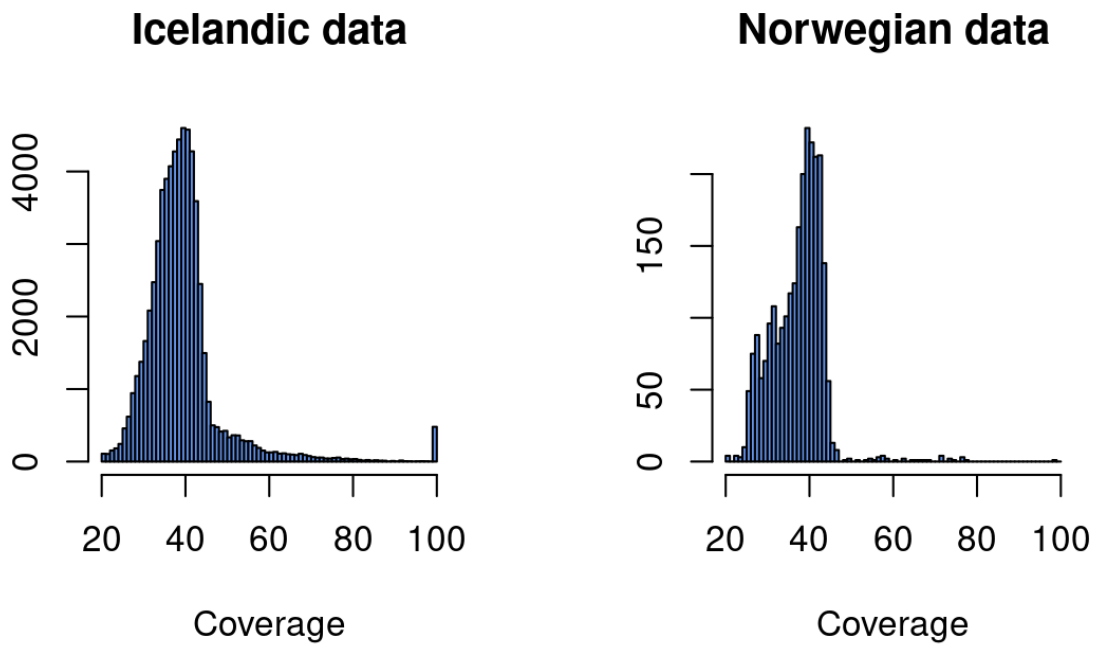
We have now improved and augmented the Genotyping Methods section as described in Response 2.3 above, by including detailed descriptions of the QC steps performed.

Q 2.6 **• Was the Norwegian study treated under the same conditions as the Icelandic population? Could you include a distributions plot of read depths between studies. Average coverage plots across chromosomes may be useful here as well.**

Response 2.6:

As described in Response 2.3 above, we now state that the WGS protocol for the Norwegian samples is exactly the same as described for the Icelandic samples. We now provide information about the average genome-wide sequencing coverage which was 37.2x (sd 6.3, min:20.1x, max:98.5x) for the Norwegian samples and 39.8x (sd 14.2, min:20.0x, max:397.8x) for the Icelandic samples.

Additionally, here below are distribution plots for the average genome-wide sequencing coverage per sample in the Icelandic and Norwegian datasets:



Q 2.7 • **Excuse my inexperience here with Graphtyper - was maximum depth (down sampling) considered? Were any other metrics used apart from AD and DP for post-variant calling QC (was hard filtering/VQSR/random forests used?) –I feel another sentence in the "Variant selection for burden test and quality control" is needed for the justification of why just AD and DP was the only metric used. For**

heterozygous count, was there a method to determine variants didn't differ by DP/2, as I couldn't imagine you would get a perfect DP/2 for most variants. This section should come before the functional prediction. The sentence about the AAscore does not provide any information, a logistic regression model on what ?? just report that variants with a AAscore > 0.8 from GraphTyper was used.

Response 2.7:

Previously we described the quality filtering that we performed additionally to the regular quality filtering we perform in general for GWAS. We do use other metrics, in particular ABHet, ABHom, QD, QUAL and PASS_ratio from GraphTyper. Maximum depth was not considered, however as we have now added to Methods section (see Response 2.3 above) only samples with a minimum genome-wide average coverage of 20x and greater were included in our study for all three study groups.

Regarding "**For heterozygous count, was there a method to determine variants didn't differ by DP/2**".

We use ABHet which is the mean allelic balance across samples, where allelic balance is calculated as the fraction of reads supporting the alternative allele against the number of reads. We excluded variants if ABHet<0.175.

In the "Variant selection for burden test and quality control" subsection in Methods, we have now moved the QC paragraph before the functional prediction paragraph, and now we just report that we use AAscore>0.8, as the reviewer suggested. The subsection is now as follows:

"Only high-quality sequence variants were considered for selection. We used the following quality metrics from GraphTyper and considered variants where ABHet > 0.175, ABHom > 0.85, QD > 6, QUAL > 10, PASS_ratio < 0.05 and AAscore > 0.8 (Supplementary Figure 11). Furthermore, to estimate the quality of the sequence variants we regressed the alternative allele counts (AD) on the depth (DP) conditioned on the genotypes (GT) reported by GraphTyper[95]. For a well-behaving sequence variant, the mean alternative allele count for a homozygous reference genotype should be 0, for a heterozygous genotype it should be DP/2 and for a homozygous alternative genotype it should be DP. Under the assumption of no sequencing or genotyping error,

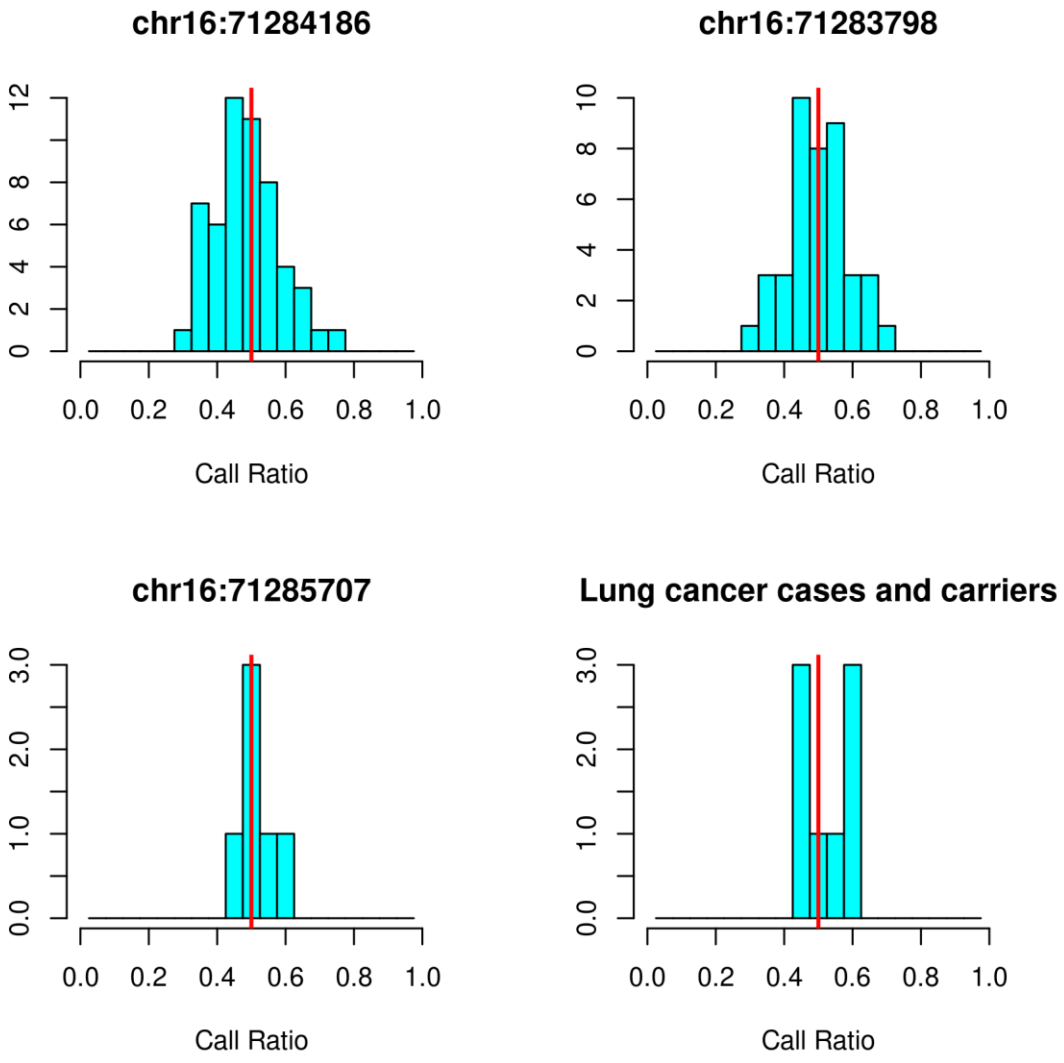
the expected value of AD should be DP conditioned on the genotype, in other words an identity line (slope 1 and intercept 0). Deviations from the identity line indicate that the sequence variant is spurious or somatic. We filter variants with slope less than 0.5 (Supplementary Figure 11).

To select the variants used to form the burden genotypes, we defined two variant selection models; (i) only LOF variants and (ii) LOF variants combined with moderate impact indels and predicted deleterious missense variants, if considered predicted deleterious by at least one of the following prediction methods: a) MetaSVM, b) MetaLR[97], using variants available in dbNSFP v4.1c[98] or c) CADD score ≥ 25[99]. In all models, we only considered variants with MAF < 2%. To include the rarest variants, we combined non-imputable variants with imputable variants."

Q 2.8 • **From my understanding, high-quality variants were selected using the mean counts to rule out somatic/artifacts calls. Since CMTR2 is a known driver – I suggest a plot showing that the variants calls are truly germline in nature. If somatic variation is observed for this gene (calls that fail the authors QC), this in my opinion is of great interest – It shows somatic change in the blood tissue outside the lung, also further supporting the authors findings in relations to little impact of smoking on this finding.**

Response 2.8:

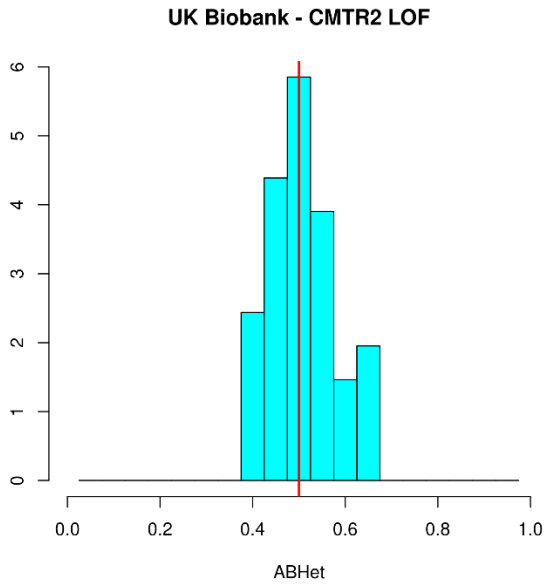
Variants that do not appear to be germline are filtered out in our quality control pipeline. One way to explore whether a variant is germline or not, is to look at the allelic-balance or call ratio distribution. Below we have plotted the distribution for the three LOF variants in *CMTR2* that are used for the burden test in Iceland:



For all three variants, the allelic-balance distribution is centered around 0.5, as is expected for germline variants. The fourth plot (bottom right) shows that the allelic-balance distribution in lung cancer patients, that are carriers of any of the three LOF *CMTR2* variants, is centered around 0.5, thereby demonstrating germline origin.

In order to inspect germline nature of the *CMTR2* LOF variants in the UK Biobank dataset, where 53 variants were combined for the *CMTR2* LOF burden, we show below the distribution for the average call ratio for these 53 variants (same as ABHet from

Graphtyper), instead of showing 53 allelic-balance distributions. The plot demonstrates that the ABHet distribution is centered around 0.5, as is expected for germline variants.



In the Icelandic and Norwegian datasets, none of the *CMTR2* LOF variants failed the QC filters applied. In the UK Biobank dataset, 9 variants failed QC, that were detected in total of 13 carriers, none of which has been diagnosed with lung cancer. Therefore, we have no evidence supporting the notion that individuals diagnosed with lung cancer have somatic *CMTR2* variants in blood tissue.

We have added these figures as Supplementary Figure 8, and this to the main text (line 278):

“Since CMTR2 has been identified as a tumor suppressor gene based on somatic mutations found in lung adenocarcinoma[45,46], we explored the call ratio distributions for the CMTR2 LOF variants in our study, which demonstrates that they are germline (Supplementary Figure 8).”

Q 2.9 • **The methods state 431,079 UKBB WGS were used, the introduction states 431,071 – which is why the number in the abstract is out by 8 compared to what the methods say.**

Response 2.9:

We have now fixed this typo. The correct number is 431,079.

Methods
Overall, my 2 biggest concerns here are related to the sex-specific effect of PPP1R15A and the lack of account for excessive relativeness/account for case-control imbalances. My recommendation would be to replicate the findings using SAIGE-GENE+ or Regenie.

Q 2.10 **• For the associations that were sex-specific cancer sites, why wasn't the burden test performed on just that one sex? If you see a ~50% reduction in developing breast cancer but 50% of the cohort is men (figure 1 – 19,121 vs 411,602 breast cancer for UK biobank, indicates men were included – male breast cancer is a different genetic disease) and the direct mechanism is acting via sex – I am unsure what this means in risk estimates and the biological context of this finding. Also why was the RNA-seq analysis not stratified by men only for the BIK finding?**

Response 2.10:

We thank the reviewer for pointing this out to us. There was inconsistency in our phenotype lists, as male cases were only included in the Icelandic data. In the UK Biobank male controls were included even though all cases were female, but in that case, performing the analysis on only women gives the same results as before because we adjusted for sex. We have fixed this now, so that we include male cases in the UK Biobank which leads to slightly different effect estimates compared to before. We note that there were no male cases in the Norwegian data. The association between the burden of LOF+missense variants in *PPP1R15A* is now OR=0.47 and P=4.4e-7 using women and men combined. For women only, the OR is 0.48 and P=1.3e-6, and for men only, the OR is 0.02 and P=0.34 (OR of 0.02 reflects the fact that none of the carriers were cancer cases).

Including males can be beneficial, since if variants are associated with breast cancer in both males and females, performing the analysis in the combined set increases

statistical power. This can be seen for example in the association results between *BRCA2* and breast cancer:

- Association of LOF burden in *BRCA2* with breast cancer combined for males and females: OR=2.1 and P=2.3e-89
- Association of LOF burden in *BRCA2* with breast cancer in females only: OR=2.0 and P=7.9e-69
- Association of LOF burden in *BRCA2* with breast cancer in males only: OR=8.6 and P=8.8e-16

But we agree with the reviewer's opinion that it is important to also report the association results in males and females separately. Therefore, we have now included the association results for males and females separately in a new Supplementary Table 4 in our manuscript. We have also added columns in Supplementary Table 1 providing the number of male and female cases analyzed.

Regarding the RNA-seq analysis for *BIK*, we see no reason to exclude women from the analysis when we are investigating the effect of the variant on transcription levels. Accordingly, when we perform our analyses separately for males and females, we do not observe sex specific difference for the effect of the genotype (p.S87G/rs757819475) on *BIK* exon 3 skipping; males (n=24 carriers, effect=3.1 SD, P = 1.4e-54) while females (n=18 carriers, effect = 3.1 SD and P = 1.7e-39).

Q 2.11 **• It was not clear to me how much of a role imputed variants played; a breakdown of this in the Supplementary Table 4 is needed. Also, a breakdown of this in the cancer data in ST1 (per cases and control status) would make it clearer. Is there a bias here because of the technology used? Would these associations go away without the use of imputed data? I think this information is critical in the interpretation of the results.**

Response 2.11:

We have now added columns in Supplementary Table 1 providing the number of cases and controls that were WGS. In Supplementary Table 6 (was ST4), we have now added a column that indicates whether a variant was imputed or not. We note that all

individuals in the UK Biobank dataset were WGS, hence no imputation was performed for those individuals.

In order to explore if any bias was generated by including imputed and WGS individuals in our burden association analyses, we performed the association analysis using only WGS individuals from all three datasets. The results from this exercise are shown in the table below. The last column provides a P-value for a heterogeneity test, comparing these effect estimates (i.e. when using WGS individuals only) with the reported estimates from the manuscript. As is expected, the P-values are slightly larger (less significant) for most burden associations, since the statistical power is lower when we reduce sample size, and for most of the results. The only difference between the effect estimates in this table and the ones in the main text of our manuscript is for the association between burden of LOF variants in *CMTR2* and lung cancer, where the association becomes stronger with lower P-value and nominally significantly higher OR when restricting to WGS individuals, possible due to better accuracy in the WGS data or because one of the more common variants included might not be a true LOF variant.

			Meta-analysis		**UK Biobank**		**Iceland**		**Norway**		
			OR	**P-value**	**OR**	**P-value**	**OR**	**P-value**	**OR**	**P-value**	**P-het**
Breast cancer	*PPP1R15A*	LOF+missense	0.47	1.4E-06	0.45	1.2E-06	0.44	1.6E-01	11.20	8.0E-02	0.92
Breast cancer	*SAMHD1*	LOF+missense	1.51	1.8E-08	1.51	2.7E-08	1.43	3.8E-01	1.08	9.5E-01	1.00
Cancers combined	*AURKB*	LOF+missense	0.82	1.5E-06	0.80	1.3E-06	0.89	1.6E-01	1.21	6.3E-01	0.65
Cancers combined	*SAMHD1*	LOF+missense	1.41	3.0E-18	1.42	5.8E-18	1.41	1.1E-01	0.24	1.3E-01	0.90
Colorectal cancer	*ATG12*	LOF	3.03	1.9E-05	2.20	7.1E-02	3.32	2.5E-04	64.70	2.8E-02	0.82
Lung cancer	*CMTR2*	LOF	5.39	9.5E-10	3.71	7.8E-05	8.08	1.3E-04	89.10	1.3E-04	0.028
Melanoma	*CMTR2*	LOF	3.63	3.3E-06	2.89	5.9E-04	3.49	1.1E-01	71.50	9.7E-05	0.87
Prostate cancer	*BIK*	LOF+missense	2.08	4.8E-11	2.24	2.4E-09	1.80	3.1E-03	0.02	4.0E-01	0.53
Prostate cancer	*SAMHD1*	LOF+missense	1.53	2.7E-07	1.57	8.0E-08	0.58	3.2E-01			0.95
Thyroid cancer	*TG*	LOF+missense	1.84	1.7E-05	1.64	3.3E-03	2.81	3.1E-04	1.14	8.5E-01	0.72

Response Table 2.11. Association results using only WGS individuals.

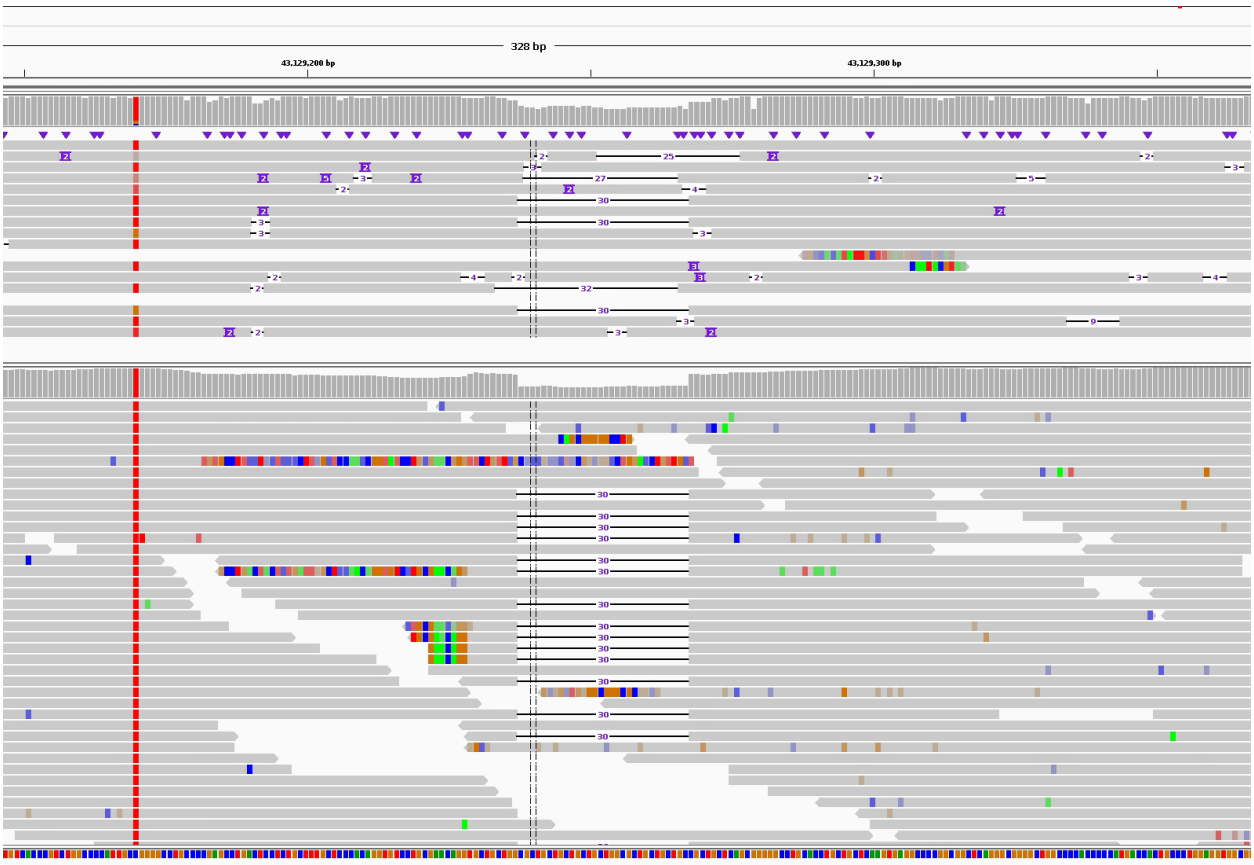
Q 2.12 • **As 55% of the variants that make up the BIK burden test come from a low copy repeat region, I would be extremely cautious in the interpretation of what this means in terms of a burden model. Could the authors make a comment regarding their justifications. Additionally, confirmation using long-read technologies (would be handy) to confirm the microsatellites are not a consequence of structural variation as there appears to be reported SVs in this region. While this finding is very interesting, more work needs to be done to unpack this.**

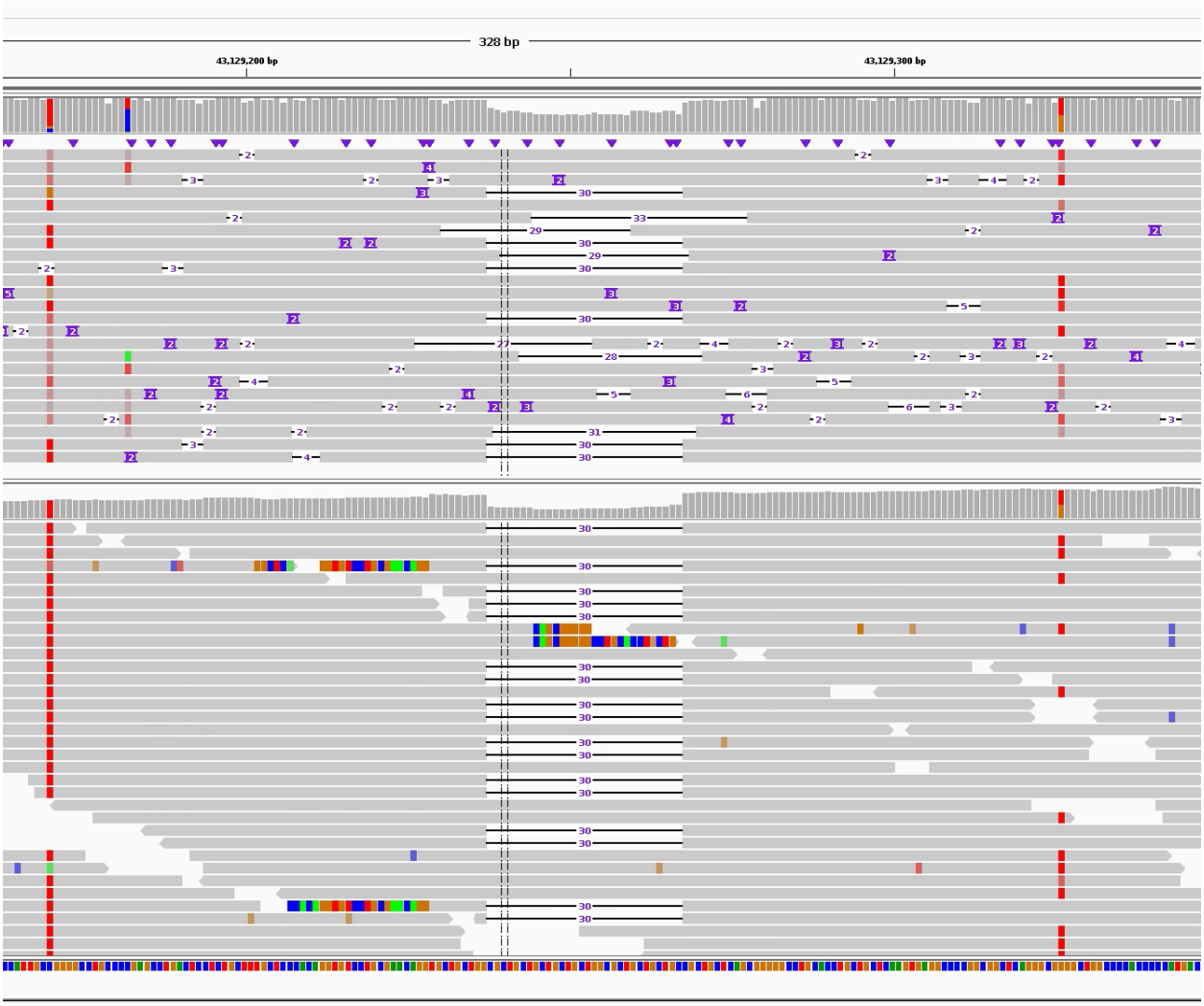
Response 2.12:

We have sequenced 9217 individuals in the Icelandic dataset with long-read sequencing using Oxford Nanopore Technologies (ONT). Of the 9217 individuals, 76 are carriers of a BIK microsatellite variant according to Illumina short-read sequencing:

- 2 are carriers of the 30 bp deletion
- 3 are carriers of the 18 bp deletion
- 71 are carriers of the 12 bp insertion

For these carriers, we used the ONT data to verify their carrier status. As an example, here below are figures showing the mapped sequencing reads of the region for two carriers of the 30bp deletion. The top section in each figure is showing ONT data and the lower section is showing Illumina data. This clearly confirms that both sequencing methods detect a 30 bp deletion.


328 bp
43,129,200 bp
43,129,300 bp



Furthermore, in the ONT sequence data from the 9217 individuals, there is no structural variant that overlaps with this region. Therefore, we can with high level of confidence confirm that the microsatellites alleles are not a consequence of a structural variant.

We agree, that some “unpacking” is yet to be done in this region, as is also true for so many other regions/loci in the human genome. However, in our opinion, the correct way forward is to make this discovery publicly known so more groups can “dig in”.

Q 2.13 • **In the tables comment (line 450) *SAMHD1 was reported while this manuscript was being prepared. Also, the sentence in the text (line 119) that mentions this. This has been identified before this manuscript has undergone review. This should be removed as the paper is focused on novel findings.**

Response 2.13:

We respectfully disagree with reviewer regarding the results for *SAMHD1* and feel that is should be included in our manuscript. We do not count *SAMHD1* as a novel finding, but we present it because we think it is of value for the original publication to have confirming results published so soon as well as demonstrating that our methods and findings are in line what other research groups are applying/publishing.

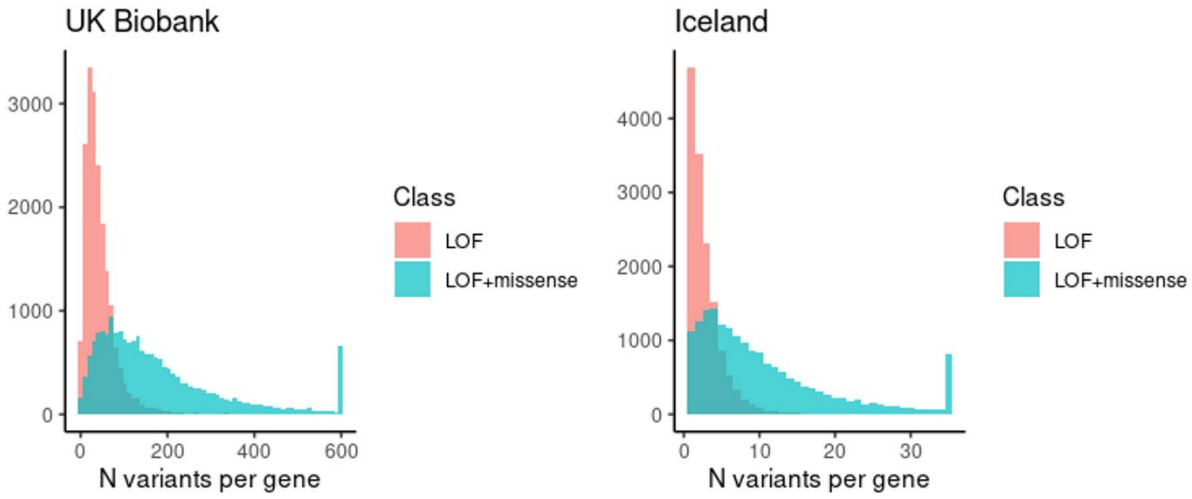
Q 2.14 • **Excessive relativeness has not been addressed and the justification for this has not been given. It appears that first- and second-degree relatives were included to boost sample size. Yet the authors have not suggested that they have correctly adjusted for this either using a relationship matrix or other statistical frameworks. While the authors have used the LD intercept to adjust the X2 statistic, this is inadequate for rare variation. Please re-run the burden tests in Regenie/SAIGE-gene+ or other methods that can handle excessive relativeness.**

Response 2.14:

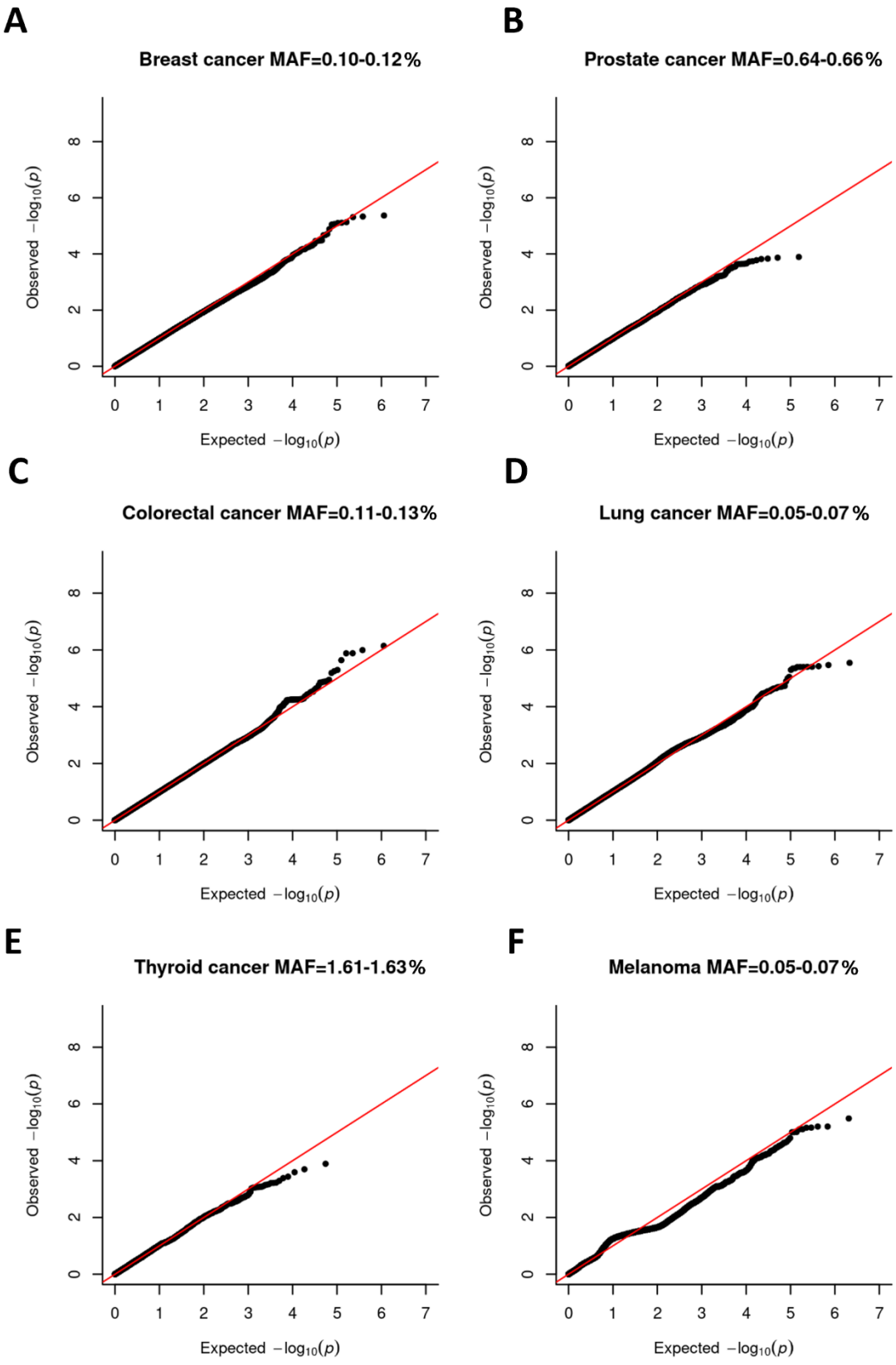
We assume that the reviewer means relatedness instead of “relativeness”.

The reviewer worries that our test statistics are inflated because of relatedness and that our novel results could therefore be false positive associations. Rare variants that are only present in one or few families, can create false positive associations. However, we note that within the burden framework we collapse rare variants together and consequently reduce the relatedness among the carriers in each test, compared to carriers of a single variant of the same frequency, as carriers of any loss of function variant in the gene are being tested.

In the UK Biobank dataset, the relatedness is much lower than in the Icelandic dataset, and in the UK Biobank, we are collapsing together substantially more variants per gene than in the Icelandic dataset (median number of variants per gene is 2 in Iceland and 34 in UK Biobank, for burden of LOF variants, and 8 in Iceland and 137 in UK Biobank, for burden of LOF+missense variants, see figure below).



To explore if the structure of our Icelandic dataset is creating false positive associations, we can utilize intronic and intergenic variants to analyze the distribution of test statistics under the null hypothesis. For the traits we are reporting novel associations for, we tested all intronic and intergenic variants in the genome with minor allele frequencies matching the frequencies of the reported burden genotypes. We found no variants associating with any of the traits (see QQ plots below), except a few variants that are correlated with known variants in *BRCA2* and *BRCA1* for breast cancer and *CDKN2A* for melanoma. We therefore removed chromosomes 13 and 17 for breast cancer and chromosome 9 for melanoma from the QQ plots. This confirms that relatedness is not producing false positive associations for these traits and frequencies, and that our association test is well calibrated, as the observed distribution of p-values matches the expected distribution under the null hypothesis.


A
Breast cancer MAF=0.10-0.12%
B
Prostate cancer MAF=0.64-0.66%
C
Colorectal cancer MAF=0.11-0.13%
D
Lung cancer MAF=0.05-0.07%
E
Thyroid cancer MAF=1.61-1.63%
F
Melanoma MAF=0.05-0.07%
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0 1 2 3 4 5 6 7
0 2 4 6 8

Regarding re-running using other method. Mbatchou et al, in the Regenie paper[3], state in the discussion that they do not recommend using Regenie in cohorts with high level of relatedness, such as founder populations, because it can become too conservative. We therefore think the Regenie method needs to be extended further before it can be used on our Icelandic dataset.

We ran the association analysis using SAIGE[4]. For technical reasons, we can only use SAIGE on our chip-typed dataset, i.e. we can not use our family imputation to gain statistical power. We therefore, ran our association test again using only chip-typed individuals and compared the results to the results from SAIGE. For the novel genes there is very little difference in the P-values, as is evident in the figures below.

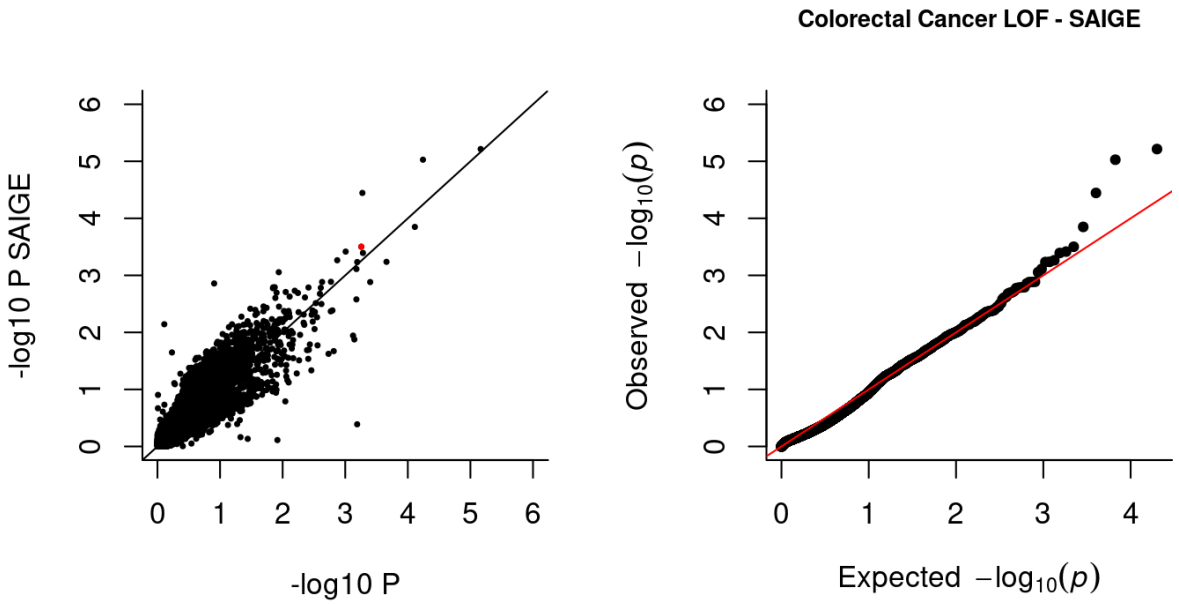


Figure for LOF+missense Colorectal Cancer results using SAIGE. Figure on left compares P-values from our association test against P-values from using SAIGE, using chip-typed individuals in Iceland. The red dot is ATG12. The figure on right is the qq plot for P-values obtained using SAIGE.

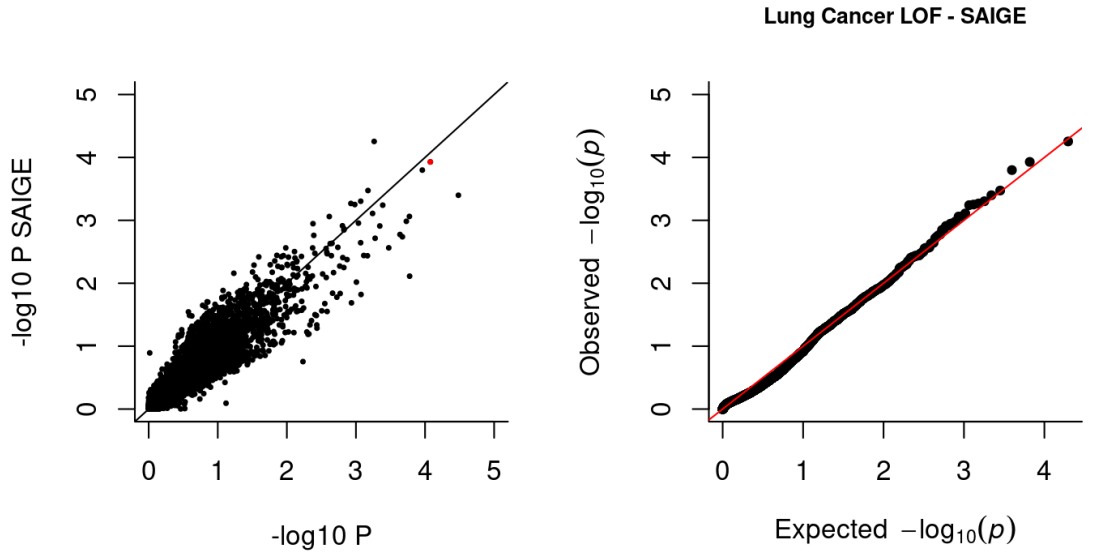


Figure for LOF+missense Lung Cancer results using SAIGE. Figure on left compares P-values from our association test against P-values from using SAIGE, using chip-typed individuals in Iceland. The red dot is CMTR2. The figure on right is the qq plot for P-values obtained using SAIGE.

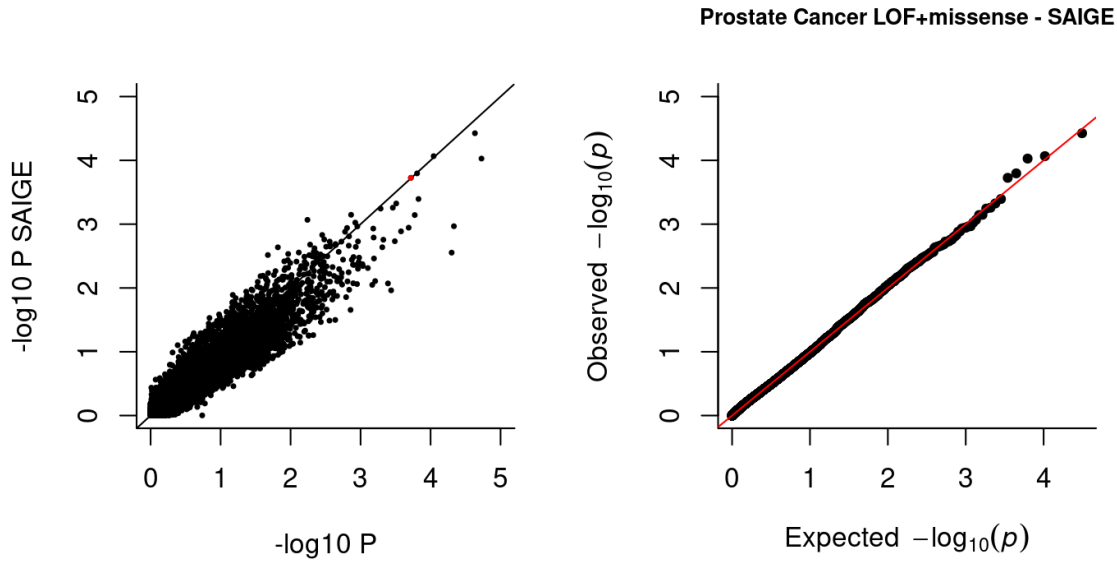


Figure for LOF+missense Prostate Cancer results using SAIGE. Figure on left compares P-values from our association test against P-values from using SAIGE, using chip-typed individuals in Iceland. The red dot is BIK. The figure on right is the qq plot for P-values obtained using SAIGE.

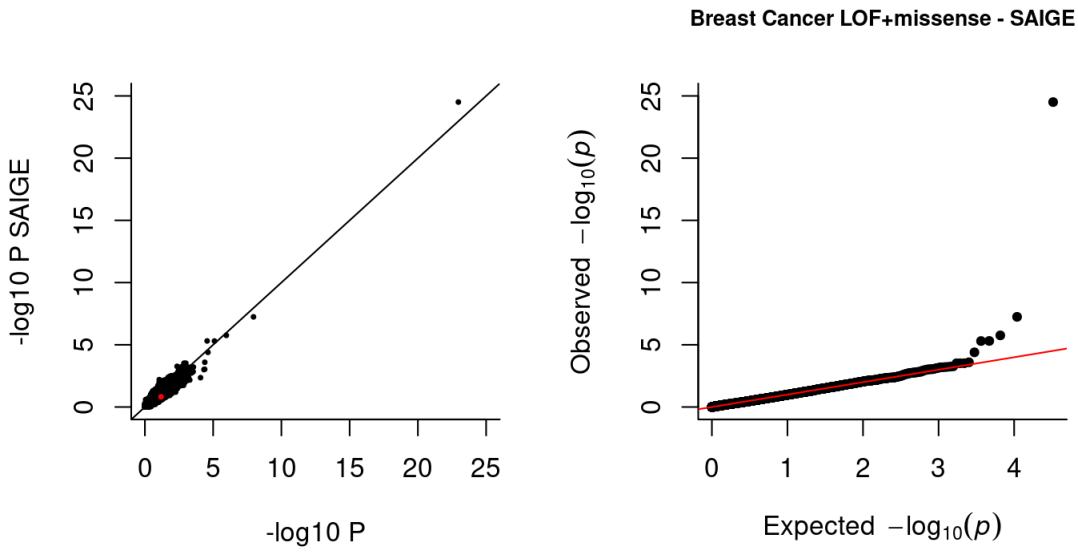


Figure for LOF+missense Breast Cancer results using SAIGE. Figure on left compares P-values from our association test against P-values from using SAIGE, using chip-typed individuals in Iceland. The red dot is PPP1R15A. The figure on right is the qq plot for P-values obtained using SAIGE.

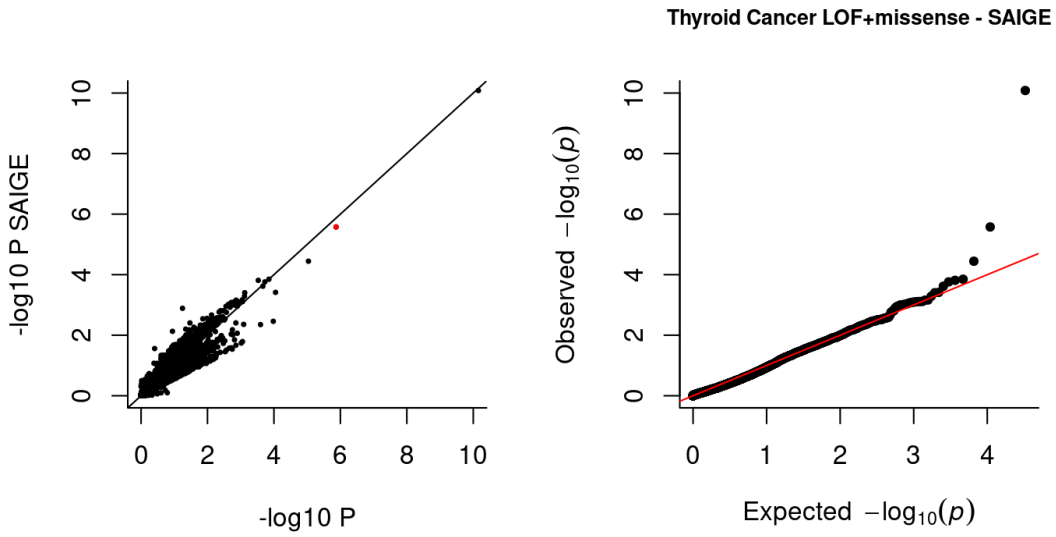


Figure for LOF+missense Thyroid Cancer results using SAIGE. Figure on left compares P-values from our association test against P-values from using SAIGE, using chip-typed individuals in Iceland. The red dot is TG. The figure on right is the qq plot for P-values obtained using SAIGE.

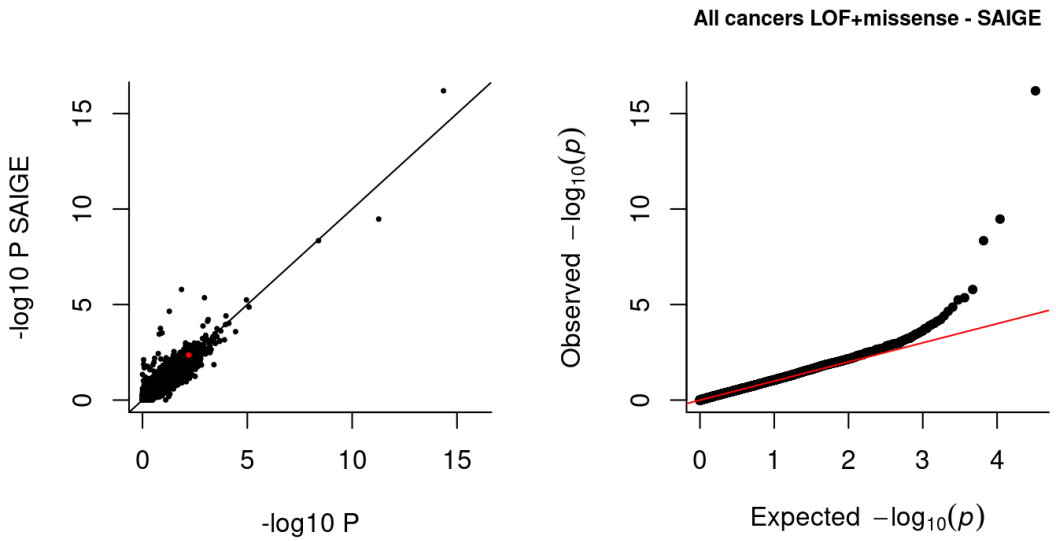


Figure for LOF+missense All Cancer results using SAIGE. Figure on left compares P-values from our association test against P-values from using SAIGE, using chip-typed individuals in Iceland. The red dot is AURKB. The figure on right is the qq plot for P-values obtained using SAIGE.

This shows that our significant association results are not due to excessive relatedness. We have added these analyses as a Supplementary Note and the figures as Supplementary Figures 16-18.

Q 2.15 • **Additional QC or justification is needed on why low confidence pLoF sites were included. For example, on splice variants (CMTR2 – 16:71289350 both T and C alleles are reported as low confidence LoFs in gnomad)**

Response 2.15:

To explore the effect of including low confidence LOF variants, we performed the associations after excluding variants that were categorized as low-confidence using LOFTEE[5]. Below are the results for the novel associations excluding low-confidence variants and the last column provides a P-value from a heterogeneity test, comparing the effect estimates with the reported estimates from the manuscript.

Response Table 2.15. Association results when removing low-confidence variants.

			Meta-analysis		Norway		UK Biobank		Iceland		
			OR	P-value	OR	P-value	OR	P-value	OR	P-value	P-het
Breast cancer	*PPP1R15A*	LOF+missense	0.45	4.3E-07	1.28	7.6E-01	0.44	3.0E-06	0.38	2.0E-02	0.92
Breast cancer	*SAMHD1*	LOF+missense	1.47	8.0E-08	1.40	3.6E-01	1.47	4.9E-07	1.47	1.0E-01	0.79
Cancers combined	*AURKB*	LOF+missense	0.84	5.2E-08	0.86	2.4E-01	0.78	7.1E-07	0.89	7.4E-03	1.00
Cancers combined	*SAMHD1*	LOF+missense	1.41	1.1E-18	0.90	7.7E-01	1.40	2.4E-16	1.56	3.3E-04	0.90
Colorectal cancer	*ATG12*	LOF	3.28	7.1E-03	14.60	7.7E-02	2.86	2.3E-02	NA	NA	0.75
Lung cancer	*CMTR2*	LOF	4.14	6.3E-10	20.30	7.3E-04	4.10	3.7E-05	3.35	2.5E-04	0.60
Melanoma	*CMTR2*	LOF	3.68	2.4E-07	11.50	2.0E-04	3.22	2.6E-04	2.55	7.2E-02	0.54
Prostate cancer	*BIK*	LOF+missense	1.86	8.8E-12	1.29	8.2E-01	2.13	1.5E-08	1.66	4.9E-05	0.87
Prostate cancer	*SAMHD1*	LOF+missense	1.53	7.7E-08	0.02	4.8E-01	1.56	8.4E-08	1.28	3.8E-01	0.95
Thyroid cancer	*TG*	LOF+missense	1.96	4.6E-11	1.14	8.4E-01	1.64	3.3E-03	2.24	9.1E-10	1.00

As the table above shows, the results are comparable to the results initially included in our manuscript, and no significant difference was observed between the initial effect estimates and the effect estimates generated after removing low confidence LOF variants. The association results for *CMTR2* and lung cancer is even stronger, although not significantly so.

Notably, the association that changed the most is *ATG12* and colorectal cancer, where the meta-OR increases slightly from 2.81 to 3.28, and the P-value increases from

P=8.0x10^{-7} to P=7.1x10^{-3}, even though the association results for UK Biobank do become more significant (from OR=2.20, P=0.071 to OR=2.86, P=0.023). This is because the only LOF variant in *ATG12* in Iceland is a start-lost variant that is categorized as a low-confidence LOF. However, we have shown, using RNA-seq data, that the variant is a true LOF variant. This shows that prediction of low confidence LOF variants is far from being unequivocal and therefore we choose to include all predicted LOF variants in our burden tests.

Q 2.16 • **Why not necessary, it would be a nice touch for authors to justify why WGS is better to perform burden testing for LoF/missense compared to WES, which in theory should be able to identify these classes of variants. As this is the largest WGS burden study to date.**

Response 2.16:

The decision to WGS a large set of samples was not made specifically for our study but was rather a part of a much larger research project and we decided to make use of the fact that these data were accessible to us. Despite the theoretical capability of whole exome sequencing (WES) to detect coding variants, WGS is preferred due to its comprehensive coverage of coding regions and improved identification of splicing and structural variants (see Halldorsson et al.[6] for more detail).

Q 2.17 • **In the stratified analysis of various smoking characteristics, could the authors provide a breakdown of carrier status for CMTR2 LoF/missense.**

Response 2.17:

We have added the number of CMTR2 LOF carriers that are cases/controls in the Ever vs Never smoker lists and the number of CMTR2 LOF carriers that have information about the number of cigarettes smoked per day in Suppl. Table 9.

Quality of presentation

Q 2.18 **Overall, the structure of the paper is fine, some parts in the methods need readjusting (mention above). The presentation of the figures is very clear. Figure 1 is presented very well. In my opinion, I think supplementary figure 3 should go in the main figures. Thank you for including QQ-plots.**

Response 2.18:

We have adjusted the methods as the reviewer has suggested (see Response 2.3 and 2.7). In our opinion, Supplementary Figure 3 does not show important additional information that is not presented in Figure 2 and we prefer to keep it in the Supplement. However, we are willing to consider other options per editorial instructions.

Appropriate use of statistics and treatment of uncertainties
This was already mentioned in the methods.

D. Conclusions:
Q 2.19 **My overall conclusion is that this paper is suited for another journal after addressing the concerns raised. The findings while interesting to a degree, I feel the authors could have done a lot more with the resource they had in relation to WGS and sample size. The robustness and validity of these findings are questionable due various issues that has been mentioned before. The reliability is reasonable, with the findings to a degree replicating in Iceland data which suggest the same areas are impacted by different variation which would make sense.**

Response 2.19:

We disagree with the reviewer when he/she states that our manuscript is not suitable for publication in Nature genetics, but eventually its fate is in the hands of the Journal's editors.

The reviewer states that our manuscript is: only interesting to a degree, feels we should have done more, finds the robustness and validity questionable but at the same time the reliability is reasonable. We find these comments both confusing and highly subjective. However, below we make an attempt to respond to them:

- Regarding the request for doing more or including more results we have already responded to a similar comment above (see Response 2.1)
- Regarding "**The robustness and validity of these findings are questionable due various issues**". We would like to point out that we perform our analysis in three different study groups, and we only report novel association findings when nominally significant in at least two of these study groups. We demonstrate the robustness of our methods and study design by confirming multiple previously published association results for high-risk cancer genes as well as discovering novel cancer risk genes. In responses to previous comments listed above, we have further demonstrated that our results are robust and valid (see Response 2.8, 2.10, 2.11, 2.12, 2.14 and 2.15).
  Furthermore, we have now extended our study to include GWAS datasets from Denmark for breast-, prostate-, colorectal-, lung cancer and melanoma (information about thyroid cancer status is not available at this timepoint). We have now added this dataset as a replication dataset to our manuscript. We note that even though this dataset is large, with around 377K samples, only few cancer cases have been WGS, and therefore it is not an ideal dataset for detection of rare variants. Nevertheless, we replicate the associations for *CMTR2* and *BIK* (see new Supplementary Table 3). For *PPP1R15A* and *ATG12* the effect sizes are in the same direction and were not significantly different from our meta-analysis results. This further demonstrates that our findings are robust and valid.

  We have now added this to the main text:

Line 108 in Introduction:

"Furthermore, a dataset including 377,448 Danes, of whom 1932 have been WGS, was used as a validation dataset."

Line 141 in Results:

“To validate the novel findings, we analyzed the Danish validation dataset, where ICD10 codes for breast, prostate, colorectal, lung cancer and melanoma were available. We replicated the associations of CMTR2 with both lung cancer and melanoma, and the association of BIK with prostate cancer (Supplementary Table 3).”

Line 170 in Results:

“In the Danish validation dataset, the effect for *PPP1R15A* was not significantly different from the meta-analysis effect (OR=0.66, P=0.26, P-het=0.35, replication power=0.53, Supplementary Table 3).”

Line 178 in Results:

“In the Danish validation dataset, we replicated this finding (OR=1.8, P=4.4×10-3, Supplementary Table 3) and meta-analyzing all four datasets, further strengthens the association between *BIK* and prostate cancer (OR=1.9, P=3.6×10-14).”

Line 239 in Results:

“In the Danish validation dataset, *ATG12* did not associate with colorectal cancer, most likely due to low replication power (OR=1.05, P=0.94, MAF=0.012%, replication power=0.35, Supplementary Table 3). Moreover, only four colorectal cancer cases in the Danish data are WGS, and consequently, we are unable to detect the rarest variants in the cases.”

Line 251 in Results:

“In the Danish validation dataset, we replicated this finding (OR=4.7, $P=2.5\times10^{-4}$, Supplementary Table 3) and meta-analyzing all four datasets, further strengthens the association (OR=4.1, $P=1.5\times10^{-12}$).”

And to Methods:

“Denmark. Danish samples and data were obtained in collaboration with the Copenhagen Hospital Biobank (CHB)[91] and the Danish Blood Donor Study (DBDS)[92]. CHB is a research biobank, which contains samples obtained during diagnostic procedures on hospitalized and outpatients in the Danish Capital Region hospitals. Data analysis within this study was performed under the CHB protocol “Developing the basis for personalized medicine in cancer – a genetic study on samples from the Danish cancer biobank (CHB-CC). CHB-CC was approved by the Capital Region Data Protection Agency (P-2022-393) and the National Scientific Committee on Health Research Ethics (NVK-2200966). The DBDS cohort is a nationwide study of ~160,000 blood donors. The Data Protection Agency (P-2019-99) and the National Committee on

Health Research Ethics (NVK-1700407) approved genetic studies within the DBDS cohort."

"The Danish genotype data was generated at deCODE by WGS 1,932 individuals and subsequent imputation into 377,448 chip-typed individuals from CHB and DBDS (described above). Long-range phasing and sequence variant calling were performed jointly for multiple cohorts of various other origin (Norwegian, N-American, Iranian, Swedish and Dutch) where 50,839 have been WGS and 1,041,174 chip-typed. The WGS protocol for the Danes was the same as described above for the Icelanders and joint variant calling was performed using GraphTyper (v.2.7.5)[96]."

F. Suggested improvements: experiments, data for possible revision

Major:

Q 2.20 • **Perform the burden test in a software package that handles excessive relativeness.**

Response 2.20:

See Response 2.14 above.

Q 2.21 • **Perform burden analysis in women only for breast cancer.**

Response 2.21:

We have now added Supplementary Table 4, showing results for breast cancer using women and men separately. See Response 2.10.

Minor:

Q 2.22 • **More details surrounding sequencing/QC/LoF filtering.**

Response 2.22:

See Response 2.3, 2.7 and 2.15.

Q 2.23 • **Is CMTR2 variants related to the somatic CMTR2 variant?**

Response 2.23:

See Response 2.8.

Q 2.24 • **Long-range sequencing/nanopore on the BIK finding**

Response 2.24:

See Response 2.12.

Q 2.25 • **Explanation of how much more information these new findings add? Do they explain more to the missing heritability?**

Response 2.25:

As these are rare variants, they will explain little of the missing heritability for each of the cancer sites. However, these findings shed new light on molecular mechanisms involved in cancer development and are clinically relevant, particularly for the individuals carrying the risk increasing variants. Adding these new genes to existing cancer gene panels would enhance the identification of high-risk individuals, thereby enabling better screening strategies and ultimately lead to earlier detection. Also see Response 2.1.

Q 2.26 • **How these findings relate to other large burden studies such as genebass.org and how the use of WGS for exonic variation is justified.**

Response 2.26:

Genebass.org only uses data from the UK Biobank. It is important to see evidence from more than one dataset to lower the probability of reporting a false positive association. Regarding WGS, see Response 2.16.

Q 2.27 **• Burden analysis of germline variants that influence non-coding elements (personally would love to see a burden of CpG sites or mQTL sites). You should see BIK again for prostate.**

Response 2.27:

Although this would be interesting to explore, performing burden tests for non-coding regions presents several challenges, particularly regarding the direction of effect. While we encourage further exploration of this topic in future studies, it is outside the scope of this paper.

Q 2.28 **G. References: appropriate credit to previous work?**
Apart from reference 91, I feel the references are appropriate.

Response 2.28:

Reference 91 is the paper “Detection of sharing by descent, long-range phasing and haplotype imputation” by Kong et al. We perform long-range phasing and we do not understand why the reviewer thinks this reference is inappropriate.

H. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

Overall, I think the lucidity is ok.

I think the clarity/context could be improved by these minor comments:
Q 2.29 **• Abstract – "is one way to shed light" is too colloquial in my opinion.**

Response 2.29:

We have changed this to "can shed light"

Q 2.30 **• 2 of these genes are not novel.**

Response 2.30:

Apart from BIK, we are not sure which other gene the reviewer is referring to. We do not count *SAMHD1* as novel. As we wrote in Response 2.1, we have clarified our interpretation of "novel" in line 137 of the manuscript:

"For six of the 34 genes, no rare coding variants had previously been reported to associate with cancer risk and we describe them as novel cancer genes: …"

Q 2.31 **• "Protection of cancer" – associate with a decreased risk of cancer?**

Response 2.31:

As suggested, we have changed this to "decreased risk of cancer".

Q 2.32 **• The number of WGS is different to what is stated in methods.**

Response 2.32:

We have fixed this typing error. In total 497,115 were WGS.

Q 2.33 **• Introduction, while the introduction is short, I feel the size is appropriate.**

Response 2.33:

We agree.

Q 2.34 • **The hypothesis of the study is a bit unclear.**

Response 2.34:

See Response 2.2 above.

Q 2.35 • **A sentence that follows up from the founder populations statement into the specifics of the study design I think is needed.**

Response 2.35:

We have now changed this part (line 99) in the Introduction to:

“Performing this type of analysis in founder populations like the Icelandic has been shown to facilitate the discovery process[12]. The founder effect allows variants under negative selection to become more common than in more outbred populations[13], potentially providing opportunity for the discovery of cancer risk variants.

To search for rare germline coding variants associating with cancer risk, we utilized three large genetic datasets of European descent…”

Reviewer #3:

Remarks to the Author:
This paper presents the results of an exome-wide association analysis for all the major cancer types, using data from the datasets (UK Biobank and datasets from Iceland and Norway). The analyses identify six "novel" associations using burden analyses based on LoF and LoF+deleterious missense variants, two of which are protective (one for all cancer).
a) **This is a very valuable analyses, such analyses have not been done for all cancers previously. While it is striking that most of associations were already known (something worthy of comment in the discussion), the identification the novel associations are interesting biologically and some could have therapeutic relevance.**
b) **The analyses follow a standard approach and are appropriate. One could argue that the Bonferoni correction is not sufficiently stringent since more than 20 cancers endpoints were considered, but this is always a somewhat arbitrary decision.**
c) **My main comment on the analysis relates to the absolute risk estimation. It is fairly meaningless derive these estimates within the cohorts since, as is pointed out, the recruitment e.g., for UK Biobank over a limited age range. If absolute risks are to be presented, it would be better to apply the estimated Ors to national incidence rates.**
d) **As per Nature Genetics guidelines, it is important that the full genome-wide set of summary statistics is deposited in a public repository, e.g., the GWAS Catalog.**

Response 3.1:

a)

As suggested by the reviewer, we now mention in the discussion (line 353) that most of the associations were already known:

"By conducting gene-based rare variant burden associations with 22 cancer sites and the combination of all cancer sites, across three datasets, from Iceland, UK and

Norway, we found 34 genes associating with cancer risk. Most of the associations are with known cancer susceptibility genes, while for 6 genes, rare coding variants have not been previously reported to associate with cancer risk."

b)

In the discovery study, we think it would be overly conservative to correct for the number of traits studied. The number of significant findings is substantial which amounts to false discovery rate (FDR) of: 0.00055 for breast cancer, 0.0035 for prostate cancer, 0.0022 for colorectal cancer, 0.000064 for lung cancer, 0.000085 for thyroid cancer, 0.012 for melanoma and 0.0027 for any cancer, at the 1.3×10^{-6} significance threshold, when assessed using the qvalue package in R. These false discovery rates, which add up to 2.1%, suggest that our multiple testing procedure is overall conservative. FDR approach using 5% discovery rate has been suggested robust in a previous GWAS publication[7].

c)

In the manuscript, we presented cumulative incidences of cancers among non-carriers and carriers for genes of interest in Figure 3 and in the main text we provided the estimated probability of being diagnosed with these relevant cancers by the age of 80. This is one way to present the observations that we made from our datasets. However, the datasets can be biased, and specifically for lung cancer and colorectal cancer we see a substantial difference in the estimated risk among non-carriers in the UK Biobank datasets compared with the Icelandic dataset.

As the reviewer suggested, we explored the national rates in Iceland and the UK.

In Iceland, we have available individual level data with year of birth and year of death, for the entire Icelandic population, and complete cancer registry data from 1955-2021.

Using this comprehensive population data, we estimated the probability of being diagnosed before the age of 80 with colorectal cancer, lung cancer, female breast cancer and prostate cancer, for individuals born between 1925-1985. The probability

estimates were; 4.2% for colorectal cancer, 5.3% for lung cancer, 14.0% for prostate cancer (among men) and 10.0% for breast cancer (among women).

These are different from the estimates of non-carriers in our chip-typed dataset, as seen in the table below. The most likely reason for the higher probability estimates for breast, prostate and colorectal cancer, is that deCODE has systematically recruited cancer cases for genetic research over the last 20 years. However, the lung cancer probability is lower, possibly because lung cancer is the most common cause of cancer death and cases are therefore more likely to die before being able to participate in our studies.

For the UK we do not have individual level data available. The Office for National Statistics has publicly available mortality rates for different ages, and the cancer research UK has released the average number of new cases per year in 2016-2018 per age groups. From this data we can estimate the probability of being diagnosed with cancer, assuming that the cancer rates and mortality rates do not change with time.

The probability estimates for being diagnosed with cancer before the age of 80 in the UK, under these assumptions are 4.9% for colorectal cancer, 5.6% for lung cancer, 13.3% for prostate cancer (among men) and 12.0% for breast cancer (among women). Comparing these numbers to our observed estimations from the UK biobank data (in table below) we again see that the UK biobank data is biased towards lower risk of lung and colorectal cancer.

	Icelandic dataset observed probability	Icelandic estimated probability for individuals born in 1925-1985	UK Biobank observed probability	UK estimated probability using national incidence rates
Breast cancer	11.1%	10.0%	11.1%	12.0%
Prostate cancer	16.6%	14.0%	14.2%	13.3%
Lung cancer	4.4%	5.3%	1.8%	5.6%
Colorectal cancer	4.3%	4.2%	1.9%	4.9%

Applying the estimated ORs to these probability estimates, the risk of being diagnosed for carriers is in the Table below:

	Icelandic estimated probability for individuals born in 1925-1985	UK estimated probability using national incidence rates
Breast cancer for carriers of *PPP1R15A*	4.7%	5.5%
Prostate cancer for carriers of *BIK*	26.7%	25.3%
Lung cancer for carriers of *CMTR2*	20.9%	23.3%
Colorectal cancer for carriers of *ATG12*	11.8%	13.8%

The estimated probability in the UK is based on incidence rates for all ethnic groups. It has been reported that the risk of diagnosis varies by ethnic groups in UK, for instance, as was reported by Lloyd et al[8], men of African ancestry were twice the risk of being diagnosed with prostate cancer compared to men of European ancestry. Since the ORs for the genes is derived from individuals of European ancestry, this can cause bias. We therefore feel it is not appropriate to report these estimates in our paper.

However, since we have the comprehensive individual level data from Iceland, we have now changed the absolute risk estimates that we before reported using the chip-typed Icelandic dataset and the UK biobank dataset, to the population derived estimates using the entire Icelandic population, and then applying the estimated OR from the meta-analysis to obtain an estimate of the risk for carriers.

Now we have changed these paragraphs in the manuscript (lines 161, 189, 236 and 265) to:

“We utilized population level data from Iceland and estimated the probability of being diagnosed with breast cancer before age 80 to be 10.0% for women born between 1925-1985, while the corresponding estimate is 4.7% (95% CI=3.4-6.2%) for carriers of LOF+missense variants in PPP1R15A.”

“We estimated the probability of being diagnosed with prostate cancer before age 80 to be 14.0% for Icelandic men born in 1925-1985. For carriers of LOF+missense variants in BIK, the estimated probability is 26.7% (95% CI=22.3-31.8%).”

“For Icelandic individuals born in 1925-1985, we estimated the probability of being diagnosed with colorectal cancer before age 80 as 4.2%, while for carriers of LOF variants in ATG12, the estimated probability is 11.8% (95% CI=7.8-17.8%).”

“We estimate the probability of being diagnosed with lung cancer before age 80 to be 5.3% for Icelandic individuals born in 1925-1985, while for carriers of LOF variants in CMTR2, the estimated probability is 20.9% (95% CI=13.4-32.7%).”

d) The summary statistics will be made available.

Other specific comments

Q 3.2

1. Introduction. It’s not clear that the founder effect argument is relevant here, given that the largest component of the data is from UKB.

Response 3.2:

One characteristic of a founder population is genetic drift, i.e. absence of some variants and a greater frequency of other variants compared to neighboring or “parental” populations. Genetic drift can also be caused by a bottleneck effect, and if it occurs within a founder population, it can potentially further exaggerate the initial frequency difference from the founder effect. In our manuscript, the reported variants in *BIK*, *ATG12*, and *CMTR2* are approximately 8 to 58-times more common in Iceland than in the UK (see Response Table 3.2 below). The greater frequency of these variants in Iceland benefits our study, as is evident in Table 2 of the main text, where the Icelandic results are a major contributor to the overall combined association results, despite the fact that the Icelandic study group is smaller than the UK-biobank study group. We mention this in the discussion, and we therefore think it is relevant to mention the founder effect in the introduction.

Response Table 3.2. Comparison of minor allele frequency in Iceland and UK.

	BIK MAF(%) rs757819475	*ATG12* MAF(%) rs770738704	*CMTR2* MAF(%) rs777823715
Iceland	0.104	0.116	0.032
UK	0.013	0.002	0.002
x-fold MAF difference	8	58	16

Q 3.3

2. The definition of likely deleterious missense variants seems reasonable. However, some justification as to why these particular tools were used – there are many other possibilities of course.

Response 3.3:

For predicting deleteriousness of missense variants, the field is rapidly evolving and there is no single universally accepted method yet. Our selection of prediction tools, namely metaSVM, metaLR, and CADD, stemmed from our prior successful experiences with these tools and them being commonly used in the literature. To not restrict ourselves to one tool, we decided to integrate results from all three methods, including variants predicted as deleterious by any of these three tools. Further, we believed it would be important to consider indels in the model, which the prediction scores do not have an evaluation for. This strategy reflects the current state of the field, but future work on developing a better prediction score would certainly be valuable.

Q 3.4

3. The SAMHD1 association has been published (or at least is in the public domain) – so it should be included among the previously described genes.

Response 3.4:

As we wrote in Response 2.13 above, we feel that the *SAMHD1* results should be included in our manuscript. We think that it is of value for the original publication to have confirming results published so soon as well as demonstrating that our methods and findings are in line what other research groups are applying/publishing.

Q 3.5

4. PPR1R15A– (I125) – over two-fold protection is an exaggeration - approximately two-fold reduction in risk would be more accurate.

Response 3.5:

We have changed “were associated with over two-fold protection or 53% lower risk” to “were associated with 53% lower risk”.

Q 3.6

5. Related to this, while the associations are quite robust, the risk estimates are much more uncertain, in particular due to the potential for winner’s curse. This should be acknowledged in the discussion.

Response 3.6:

As we wrote in Response 2.19, we have now added a replication dataset to our study.
We have added this sentence to the discussion (line 441):

“We note that as in all genetic discovery studies, there is a possibility of a winner’s curse, which can lead to inflated effect estimates. We replicated our findings for *CMTR2* and *BIK* in a Danish dataset with comparable effect estimates. However, further replication studies are essential to obtain more accurate risk estimations for all the novel genes we describe here.”

Q 3.7

6. It appears that both incident and prevalent cases were included in the

analyses, please clarify in the methods. Were only cancers identified through linkage to cancer registries included or were self-reported cancers also included?

Response 3.7:

We do not include self-reported cancer diagnosis in our study.

For UK Biobank, data on cancer diagnosis up until 2022 were used in our study and provided to the UK Biobank by the Medical Research Information Service based at the NHS Information Center and The Information Services Division, a part of the NHS Scotland. For Iceland we used cancer diagnosis from the Icelandic Cancer Registry up until 2022.

While for Norway, as is stated in Methods, the cancer data were derived from a biobank of blood samples collected over a period of 25 years as a patient self-referral initiative for overrepresentation of cancer in families, with both clinical and research intent.

We now make this clearer in Methods, by adding that for both Iceland and UK Biobank cancer diagnosis up until 2022 we used for the analysis (yellow highlights):

"Iceland. The Icelandic cancer datasets have been previously described[74,81]. The information on cancer diagnosis is obtained from the Icelandic Cancer Registry (ICR) and is based on ICD codes and histology codes (systematized nomenclature of medicine, SNOMED), and in this study we utilize registered diagnosis before 2022. The ICR has registered all solid, non-cutaneous cancer in the Icelandic population since 1955[82], while the registration of hematological and skin cancers started in the 1980s. The smoking dataset has been described previously[82]. The study was approved by the Icelandic Data Protection Authority and the National Bioethics Committee (VSN-18-115, VSN-18-029, VSN-17-204, VSN-20-004, VSN-17-137, VSN-17-138). All genotyped participants signed a written informed consent.

UK Biobank. The UK Biobank study is a large prospective cohort study of around 500,000 individuals, who enrolled in the study between 2006 and 2010 throughout the

UK and were aged 40-69 years at recruitment[84]. Extensive phenotypic information has been collected for the participants from hospital records and general practitioners. Data on cancer diagnoses, date of diagnosis (up until 2022) and deaths were provided to the UK Biobank by the Medical Research Information Service based at the NHS Information Center and The Information Services Division, a part of the NHS Scotland. The UK Biobank data were obtained under application number 56270. All participants provided written informed consent and The North West Research Ethics Committee reviewed and approved the UK Biobank protocol (ref. 06/MRE08/65)."

Q 3.8

7. BIK (I164) what does "driven" mean here? What happens if p.R618H is excluded. More generally, is there evidence for any of the reported genes that the effect size varies by variant?

Response 3.8:

"Driven by" means here that the strongest individual variant that was used for the burden test is p.S87G (we assume the reviewer meant to write p.S87G, not p.R618H which is a variant in *PPP1R15A*). The association between p.S87G and prostate cancer in the Icelandic data has OR=2.8 and $P=1.4x10^{-4}$, while the burden association has OR=1.7 and $P=4.9x10^{-5}$. If we exclude p.S87G, the association is OR=1.5 and $P=5.8x10^{-3}$. Even though p.S87G is driving the *BIK* burden association in Iceland, *BIK* would still be significant in the meta-analysis if p.S87G would be excluded.

For clarity we have changed this:

"In the Icelandic dataset, the association between *BIK* and prostate cancer was driven by p.S87G… "

To this:

“In the Icelandic dataset, the variant in *BIK* that is most significantly associated with prostate cancer is p.S87G…”

And similar changes were made in lines 204 and 269 in the main text.

In general, for most burden associations, some variants are driving the associations, i.e. associating more strongly than other variants that were included, and one could say that the included variants that are not associating are “diluting” the burden association results. We have a paragraph about this in the discussion, where we talk about the BIK association as an example:

“The rationale for performing a burden test is to boost statistical power to detect associations between rare LOF variants and phenotypes, assuming that the selected variants all have effects in the same direction. This assumption does not always hold, and most often we are including some variants that are not truly causing loss of the gene’s function, thereby resulting in a “diluted” association signal reflected by lower effect estimates. Including missense variants can result in this dilution, as is evident in Figure 2, where all effect estimates for burden of LOF+missense variants and colorectal cancer are substantially lower than for burden of LOF variants only. On the other hand, including missense variants allowed us to observe cancer associations for PPP1R15A, BIK, TG and AURKB, that would otherwise not have been detected due to lack of statistical power in the associations for burden of LOF variants. Annotated LOF variants can also dilute the signal and for instance, the association between burden of LOF variants in BRCA2 and breast cancer, has an OR of 2.1, while the OR for the well-known 999del5 variant in Iceland[77] is 16.4. This difference in effect estimates is due to the most common BRCA2 LOF variant being K3326*, which does not confer risk of breast cancer[74]. Another dilution example is the LOF+missense burden association between BIK and prostate cancer, where the burden OR is 1.9, whereas for both the p.S87G variant and the 30 base-pair deletion, the OR is 2.7. These examples highlight the advantages of using burden tests for gain in statistical power, as well as the importance of further scrutinizing the single variant associations of the variants included in the burden test.”

Q 3.9

8. Phenome-wide associations (l311). This does not seem quite the right analysis, because the power for different cancers varies hugely. I think you need

to do a test of heterogeneity of the OR among cancer types, and/or evaluate the association for all other cancers combined, after the identified associations are removed.

Response 3.9:

We agree with the reviewer that the power for detecting associations with other cancer sites varies and is low for rarer cancer sites. Before we wrote that we found *PPP1R15A*, *BIK*, *ATG12* and *TG* associated specifically with breast-, prostate-, colorectal- and thyroid cancer, respectively, but since we might not have the power to detect association with other cancer sites, we have now added a Supplementary Table 15 that reports the P value for a test of heterogeneity between the effect sizes and changed the text to:

"Correcting for multiple testing, we found that *PPP1R15A*, *BIK*, *ATG12* and *TG* did not significantly associate with other cancer sites than breast-, prostate-, colorectal- and thyroid cancer, respectively ($P \geq 0.05/25 = 2.0 \times 10^{-3}$). *CMTR2* associated with both lung cancer and cutaneous melanoma, and associated suggestively with cervical cancer (OR=4.4, P=0.0076)."

References

1. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–860 (2015).
2. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).
3. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).
4. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335–1341 (2018).
5. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
6. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
7. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet* **49**, 1385–1391 (2017).
8. Lloyd, T. *et al.* Lifetime risk of being diagnosed with, or dying from, prostate cancer by major ethnic group in England 2008-2010. *BMC Med* **13**, 171 (2015).

Decision Letter, first revision:

13th Mar 2024

Dear Dr Ivarsdottir,

First, please accept my apologies for the delay in returning this decision to you. Thank you for bearing with me.

Your Article, "Gene-based burden test yields rare germline variants in six novel cancer genes" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

We therefore invite you to revise your manuscript taking into account all reviewer comments. To the point about creating an online tool, we agree that increasing the accessibility of data would be

beneficial and we would therefore encourage you to think about ways to facilitate this, whether this is through the development of an online browser or through other means. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more

information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Safia Danovi, PhD
Senior Editor, Nature Genetics
ORCiD: 0009-0007-7822-5479

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
I thank the authors for responding to my comments and those of others. From my perspective outstanding issues are:
The authors rebut my assertion that the data should be made available in a more accessible form by stating that they do not have the resources to design and maintain an online tool to present results is rather limp. Decode is far better funded to put together such a tool which could easily be modelled on what is already out there from other research groups.
In response to other reviewer comments, they now go into some detail about low confidence genotypes. Clearly the new data concerns over the association of ATG12 with colorectal cancer, since this is entirely driven by the Icelandic. While the authors they state that they have confidence in the LOF genotypes given some RNA sequencing data without support from other cohorts this association is at best tenuous. Moreover, this association brings up the issue the potential impact of relatedness on study findings.

Reviewer #2:
Remarks to the Author:
The authors have addressed my concerns with this manuscript.

I have no further comments on this manuscript.

Congratulations on the findings in this manuscript, as I sit here at the airport in the layover, having to travel across the world to be with my sister that just yesterday got diagnosed with metastatic breast cancer with weeks to live - I hope these findings will contribute to one day stopping this terrible disease taking lives.

Reviewer #3:
Remarks to the Author:

The authors have addressed many of my concerns, but I still have concerns particularly about the cumulative risk estimates, as well as a couple of more specific points below:

Response 3.1

As previously noted, the estimated risks based directly on the studies used in the analysis are not particularly meaningful as they have limited age ranges and (as also noted by the authors in the case of the Icelandic data) have been differentially sampling with respect to cancer. I am pleased to see that the paper now includes estimates based on applying the estimated ORs to population rates. There are a couple of points to address however:

- The methods should contain a description of how the cumulative risks were derived.

- If using Iceland as the base population, it would be better to use rates from 1955 onwards, when there was reasonably complete cancer registration (arguably better to use more recent rates, since the interest is more in predicting future risks than risks in the past).

- It appears that Figure 3 is still based on the cumulative risks in the cohort. As noted, this is misleading (the Figure refers to "cumulative risk in Iceland"). The figures should either be removed, or replaced by estimated cumulative risks using population rates, consistent with the text (ideally with confidence limits).

- The authors note that the main analyses were based on individuals of European ancestry (in UK Biobank) and that cumulative rates based on UK rates would not necessarily be appropriate for non-Europeans. That argument presumably also applies in Iceland, though to a lesser extent. Please clarify in the methods.

I am pleased to see that the full set of summary statistics from these analyses will be released into the public domain.

Response 3.4

Please also reference Wilcox et al (also noted by reviewer #1).

Response 3.7

Please indicate in the methods whether both prevalent (i.e. cases prior to recruitment) and incident cancers in UK Biobank were included as cases (I presume this is the case but please be explicit).

Author Rebuttal, first revision:

Round 2 of Response to Reviewers

Reviewer #1:
Remarks to the Author:

I thank the authors for responding to my comments and those of others. From my perspective outstanding issues are:
The authors rebut my assertion that the data should be made available in a more accessible form by stating that they do not have the resources to design and maintain an online tool to present results is rather limp. Decode is far better funded to put together such a tool which could easily be modelled on what is already out there from other research groups.

Response

To increase the accessibility of data, we now provide an excel file as a Supplementary Table 17, with association results from the meta-analysis for all 23 cancer phenotypes and all genes tested. Then the results can easily be looked up per gene and phenotype.

In response to other reviewer comments, they now go into some detail about low confidence genotypes. Clearly the new data concerns over the association of ATG12 with colorectal cancer, since this is entirely driven by the Icelandic. While the authors they state that they have confidence in the LOF genotypes given some RNA sequencing data without support from other cohorts this association is at best tenuous. Moreover, this association brings up the issue the potential impact of relatedness on study findings.

Response

In this comment, Reviewer 1 is commenting on a response to a comment from another reviewer and apparently misunderstanding the content of it. To clarify this misunderstanding, we want to stress the fact that we do not make use of any low confidence genotypes in our study, we only consider high-quality genotypes for our analyses. However, the start-lost variant in *ATG12*, reported in our paper, is predicted by LOFTEE (Loss-Of-Function Transcript Effect Estimator), using bioinformatical methods, to be a low-confidence LOF variant. This prediction only indicates that there is some uncertainty whether the variant causes a loss of function of the protein. We, on the other hand, analyze our RNAseq data, and see that the variant is associated with lower expression of *ATG12*, thereby supporting the notion that it is indeed a LOF variant. As a further clarification, we would like to point out that in the LOFTEE paper, the authors state that the filtering strategy of LOFTEE is conservative in the interest of increasing specificity and variants that are categorized as low-confidence can be true LOF variants[1].

Furthermore, we would like to point out that the *ATG12* results are not entirely driven by Icelandic results, since they do replicate in the Norwegian data. Also, the variant has a consistent association effect in the UK Biobank, despite being much rarer in the non-Icelandic datasets. Additionally, by running SAIGE on the Icelandic dataset, we have already demonstrated in our previous response to reviewers' comments that the association is not due to relatedness. Finally, in our paper, we do not report any association results that are nominally significant in only one study group, not even if they surpass our genome-wide significance threshold.

Reviewer #2:
Remarks to the Author:
The authors have addressed my concerns with this manuscript.

I have no further comments on this manuscript.

Congratulations on the findings in this manuscript, as I sit here at the airport in the layover, having to travel across the world to be with my sister that just yesterday got diagnosed with metastatic breast cancer with weeks to live - I hope these findings will contribute to one day stopping this terrible disease taking lives.

Response
We thank you for the kind words about our research and we are very sorry to hear about the difficult circumstances you are facing.

Reviewer #3:
Remarks to the Author:
The authors have addressed many of my concerns, but I still have concerns particularly about the cumulative risk estimates, as well as a couple of more specific points below:

Response 3.1

As previously noted, the estimated risks based directly on the studies used in the analysis are not particularly meaningful as they have limited age ranges and (as also noted by the authors in the case of the Icelandic data) have been differentially sampling with respect to cancer. I am pleased to see that the paper now includes estimates based on applying the estimated ORs to population rates. There are a couple of points to address however:

- The methods should contain a description of how the cumulative risks were derived.

Response
We have now added this to the methods section:

"The cumulative risk of being diagnosed with specific types of cancers in the Icelandic population was computed for all individuals in our Icelandic genealogical database that were born in Iceland between 1935 and 1985 (n=234K). We note that 135K (58%) of them have been chip-typed. For each cancer type, the time variable t was defined as the age at cancer diagnosis or follow-up time in years for those without diagnosis. The probability of being diagnosed before age 80 (t=80) was computed as $P(T \leq t)=1-S(t)$, where S(t) is the survival function. We utilized the Kaplan-Meier estimator with the 'survfit' function in R, from the 'survival' package, to compute S(t). To estimate the cumulative risk for carriers, we applied the estimated OR from the meta-analysis to the population derived estimates."

- If using Iceland as the base population, it would be better to use rates from 1955 onwards, when there was reasonably complete cancer registration (arguably better to use more recent rates, since the interest is more in predicting future risks than risks in the past).

Response

We did not use reported rates for Iceland. As we now describe in Methods, we computed the cumulative incidence for all individuals in our genealogical database who were born in Iceland using cancer diagnosis obtained from the Icelandic cancer registry after 1955. We only include individuals born after 1935, so that the individuals were not older than 20 years old when the ICR started registering all cancer diagnoses.

- It appears that Figure 3 is still based on the cumulative risks in the cohort. As noted, this is misleading (the Figure refers to "cumulative risk in Iceland"). The figures should either be removed, or replaced by estimated cumulative risks using population rates, consistent with the text (ideally with confidence limits).

Response

We agree with the reviewer that the y-axis labels in Figure 3 were misleading and have now changed it from "Cumulative incidence of [breast/prostate/colorectal/lung] cancer in Iceland" to "Cumulative incidence of [breast/prostate/colorectal/lung] cancer in the Icelandic chip-typed dataset".

These figures illustrate the observed cumulative incidences for carriers and non-carriers of the burden genotypes in the datasets, highlighting the difference between these groups. It is important to note that no dataset that is based on voluntary participation will perfectly represent the entire population that it is drawn from, and as we noted before, the chip-typed dataset seems to be slightly biased with an overrepresentation of cancer cases for breast and prostate cancer and underrepresentation of lung cancer cases. We do not think it is optimal to plot the population cumulative incidence in the figures, as we do not know the carrier status in that part of the data, and the aim of the figures are to display the difference between carriers and non-carries.

In response to the reviewer's concern, we have added scaled y-axes on the right of Figures 3.A, C, E and G, to address the observed bias. Specifically, the right y-axes are scaled by the ratio of the cumulative incidence at age 80 for all individuals born in Iceland between 1935-1985 to the cumulative incidence at age 80 in the chip-typed dataset.

Updated Figure 3 and the corresponding legend is displayed here:

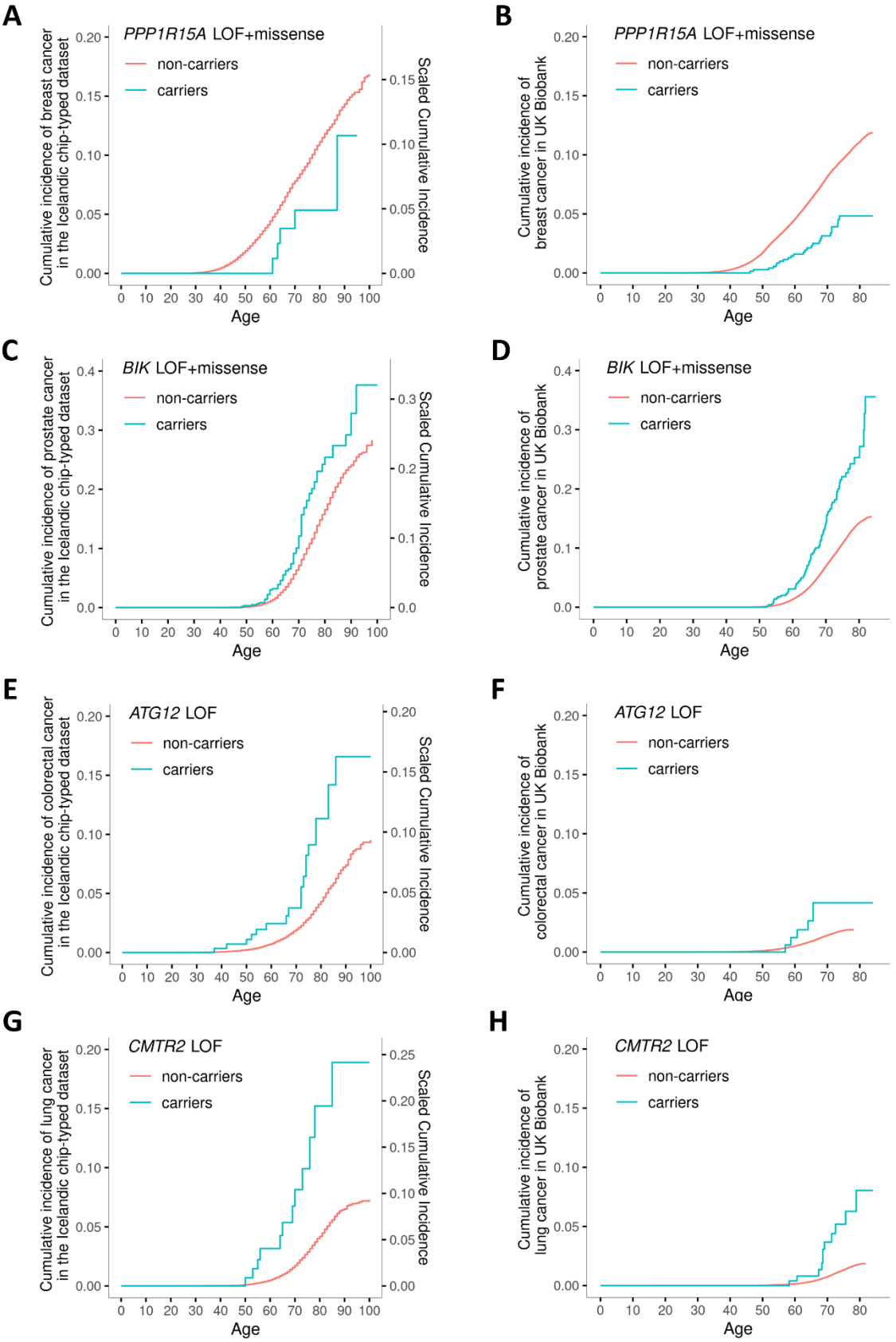

A
PPP1R15A LOF+missense
non-carriers
carriers
Cumulative incidence of breast cancer in the Icelandic chip-typed dataset
Scaled Cumulative Incidence
Age
B
PPP1R15A LOF+missense
non-carriers
carriers
Cumulative incidence of breast cancer in UK Biobank
Age
C
BIK LOF+missense
non-carriers
carriers
Cumulative incidence of prostate cancer in the Icelandic chip-typed dataset
Scaled Cumulative Incidence
Age
D
BIK LOF+missense
non-carriers
carriers
Cumulative incidence of prostate cancer in UK Biobank
Age
E
ATG12 LOF
non-carriers
carriers
Cumulative incidence of colorectal cancer in the Icelandic chip-typed dataset
Scaled Cumulative Incidence
Age
F
ATG12 LOF
non-carriers
carriers
Cumulative incidence of colorectal cancer in UK Biobank
Age
G
CMTR2 LOF
non-carriers
carriers
Cumulative incidence of lung cancer in the Icelandic chip-typed dataset
Scaled Cumulative Incidence
Age
H
CMTR2 LOF
non-carriers
carriers
Cumulative incidence of lung cancer in UK Biobank
Age

Figure 3. The cumulative incidence of cancer among carriers and non-carriers. The cumulative incidence is shown for **A-B.** breast cancer among carriers and non-carriers of LOF+missense variants in *PPP1R15A* in the Icelandic chip-typed dataset and the UK Biobank, **C-D.** prostate cancer among carriers and non-carriers of LOF+missense variants in *BIK* in the Icelandic chip-typed dataset and the UK Biobank, **E-F.** colorectal cancer among carriers and non-carriers of LOF variants in *ATG12* in the Icelandic chip-typed dataset and the UK Biobank and **G-H.** lung cancer among carriers and non-carriers of LOF variants in *CMTR2* in the Icelandic chip-typed dataset and the UK Biobank. In A, C, E, G the cumulative incidence is shown among chip-typed individuals in the Icelandic dataset (N=173,025), and the left y-axis is scaled by the ratio of cumulative incidence at age 80 for all individuals born in Iceland between 1935-1985 (Methods) to the cumulative incidence at age 80 in the chip-typed dataset. The ratios are 0.91 in A, 0.85 in C, 0.98 in E and 1.28 in G. In B, D, F, H the cumulative incidence is shown among white British/Irish individuals in the UK Biobank (N=431,079).

- The authors note that the main analyses were based on individuals of European ancestry (in UK Biobank) and that cumulative rates based on UK rates would not necessarily be appropriate for non-Europeans. That argument presumably also applies in Iceland, though to a lesser extent. Please clarify in the methods.

Response

As clarified above, we are not using reported incidence rates from Iceland. We used all 234K individuals in our Icelandic genealogical database born in Iceland between 1935-1985 to estimate the cumulative risk. Of the 234K individuals, we have chip-typed 135K individuals (58%) and of those ~0.06% have <90% European ancestry (analyzed using ADMIXTURE[2]). Therefore, we assume the argument does not apply to the Icelandic data.

I am pleased to see that the full set of summary statistics from these analyses will be released into the public domain.

Response 3.4

Please also reference Wilcox et al (also noted by reviewer #1).

Response
We have now also referenced Wilcox et al.

Response 3.7

Please indicate in the methods whether both prevalent (i.e. cases prior to recruitment) and incident cancers in UK Biobank were included as cases (I presume this is the case but please be explicit).

Response
We have added this sentence to methods:
“Both prevalent and incident cancer diagnosis (up until 2022) were included as cases.”

References

1. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
2. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).

Decision Letter, second revision:

27th Jun 2024

Dear Dr Ivarsdottir,

First, please accept my sincere apologies for the delay in returning this decision to you.

Your Article, "Gene-based burden test yields rare germline variants in six novel cancer genes" has now been seen by 2 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

We therefore invite you to revise your manuscript taking into account all comments by Reviewer #3. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available
here.
Refer also to any guidelines provided in this letter.

*3) Include a revised version of any required Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.
A revised checklist is essential for re-review of the paper.

Please be aware of our guidelines on digital image standards.

Please use the link below to submit your revised manuscript and related files:

[redacted]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Safia Danovi, PhD
Senior Editor, Nature Genetics
ORCiD: 0009-0007-7822-5479

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The authors have addressed my concerns.

Reviewer #3:
Remarks to the Author:
The paper now includes a clearer description of the methods used to compute the cumulative risks and estimates, but there is a couple of points that should still be clarified in the methods, and I note a possible error in Figure 3 that may need correcting.

The methods state: "To estimate the cumulative risk for carriers, we applied the estimated OR from the meta-analysis to the population derived estimates."

This doesn't seem to correspond with what was done in Figure 3 however. It looks like figures in Figure 3 are Kaplan-Meier curves based on the genotyped individuals. (With steps at the cancer diagnoses: if you applied estimate ORs to the population rates you would get much smooth curves in both carriers and non-carriers). The figures also provide scaled cumulative risks based on computing a ratio of the observed cumulative risks in the genotyped cohort to the Icelandic incidence rates (according to the figure legend). If this is correct, the methods should be corrected so that they correspond with the description in the figure legend.

In the methods, it is worth noting that the reason for the adjustment is to allow for potential differential sampling with respect to cancer diagnoses in the genotyped cohort (hence, oversampling of breast and prostate cancer and underrepresentation of lung cancer, as noted in the rebuttal, which makes sense).

On a minor point, if the adjustments are to be made this way it would be more natural to compute the ratio of the cumulative rates (sum of the age-specific incidences) rather than the cumulative risks (i.e. 1-exp(-cumulative rates)) . These would then correspond to the relative risk in the cohort compared to the population (assumed constant over age). The difference would be minor while the cumulative risk is low, but while correcting the methods paragraph, please clarify what the ratios refer to.

Figure 3. In figure 3, the scaled estimates are actually higher for breast and prostate cancer and lower of lung cancer (for example in G the lifetime lung cancer risk in carriers looks to be ~18% unadjusted and 14% adjusted). However the ratios given in the figure legend imply the opposite (e.g. 1.28 for lung cancer, as also noted in the rebuttal). It looks to me that the adjustments have been inverted – please check and if necessary correct the figures.

The other points have been satisfactorily addressed.

Author Rebuttal, second revision:

Reviewer #3:

Remarks to the Author:
The paper now includes a clearer description of the methods used to compute the cumulative risks and estimates, but there is a couple of points that should still be clarified in the methods, and I note a possible error in Figure 3 that may need correcting.

The methods state: "To estimate the cumulative risk for carriers, we applied the estimated OR from the meta-analysis to the population derived estimates."

This doesn't seem to correspond with what was done in Figure 3 however. It looks like figures in Figure 3 are Kaplan-Meier curves based on the genotyped individuals. (With steps at the cancer diagnoses: if you applied estimate ORs to the population rates you would get much smooth curves in both carriers and non-carriers). The figures also provide scaled cumulative risks based on computing a ratio of the observed cumulative risks in the genotyped cohort to the Icelandic incidence rates (according to the figure legend). If this is correct, the methods should be corrected so that they correspond with the description in the figure legend.

In the methods, it is worth noting that the reason for the adjustment is to allow for potential differential sampling with respect to cancer diagnoses in the genotyped cohort (hence, oversampling of breast and prostate cancer and underrepresentation of lung cancer, as noted in the rebuttal, which makes sense).

On a minor point, if the adjustments are to be made this way it would be more natural to compute the ratio of the cumulative rates (sum of the age-specific incidences) rather than the cumulative risks (i.e. 1-exp(-cumulative rates)) . These would then correspond to the relative risk in the cohort compared to the population (assumed constant over age). The difference would be minor while the cumulative risk is low, but while correcting the methods paragraph, please clarify what the ratios refer to.

Figure 3. In figure 3, the scaled estimates are actually higher for breast and prostate cancer and lower of lung cancer (for example in G the lifetime lung cancer risk in carriers looks to be ~18% unadjusted and 14% adjusted). However the ratios given in the figure legend imply the opposite (e.g. 1.28 for lung cancer, as also noted in the rebuttal). It looks to me that the adjustments have been inverted – please check and if necessary correct the figures.

The other points have been satisfactorily addressed.

Response

Regarding the error in Figure 3, an incorrect version of the figure was uploaded in the main text file, which had inverted ratios as the reviewer correctly identified. The correct version of the figure, which was previously included in the Response to Reviewers file, has now been uploaded to the main text file.

Regarding the other comments, we want to first outline our analysis.

We have two Icelandic datasets where we compute the cumulative incidence of cancer: i) the population level dataset with all individuals born in Iceland between 1925-1985 (N=234K), and ii) the chip-typed dataset (N=173K). We note that 135K (135K/234K=58%) of the individuals in the population level dataset have been chip-typed.

Supplementary Table 17:

		i) Icelandic population level dataset Individuals born in Iceland between 1925-1985 (N=234K)		ii) Icelandic chip-typed dataset (N=173K)		
Cancer	Burden Genotype	(a) Estimated probability of being diagnosed before age 80	(b) Estimated probability of being diagnosed before age 80 for carriers	(c) Estimated probability of being diagnosed before age 80 for non-carriers	(d) Estimated probability of being diagnosed before age 80 for carriers	Scaling ratio (a/c)
Breast cancer	*PPP1R15A* LOF+missense	10.20%	4.80%	11.10%	5.30%	0.91
Prostate cancer	*BIK* LOF+missense	14.10%	26.80%	16.60%	25.40%	0.85
Lung cancer	*CMTR2* LOF	5.40%	21.60%	4.20%	15.20%	1.28
Colorectal cancer	*ATG12* LOF	4.20%	11.80%	4.30%	11.30%	0.98

In the population level dataset, we estimate the probability of being diagnosed at age 80 with $P(T \leq t)=1-S(t)$, where $t=80$ and $S(t)$ is the survival function estimated with the Kaplan-Meier method. The population derived estimates are in column **a** in the table above. To estimate the corresponding probabilities for carriers in the Icelandic population (where carrier status is not

known for all individuals) we applied the estimated OR from the meta-analysis to the population derived estimates, results shown in column **b**. These estimates are reported in the main text of the paper:

"We utilized population level data from Iceland and estimated the probability of being diagnosed with breast cancer before age 80 to be 10.2% for women born in Iceland between 1925-1985, while the corresponding estimate was 4.8% (95% CI=3.4-6.2%) for carriers of LOF+missense variants in *PPP1R15A*."

"We estimated the probability of being diagnosed with prostate cancer before age 80 to be 14.1% for men born in Iceland between 1925-1985. For carriers of LOF+missense variants in *BIK*, the estimated probability is 26.8% (95% CI=22.3-31.8%)."

"For Icelandic individuals born in Iceland in 1925-1985, we estimated the probability of being diagnosed with colorectal cancer before age 80 as 4.2%, while for carriers of LOF variants in *ATG12*, the estimated probability was 11.8% (95% CI=7.8-17.8%)."

"We estimated the probability of being diagnosed with lung cancer before age 80 to be 5.4% for Icelandic individuals born in Iceland in 1925-1985, while for carriers of LOF variants in CMTR2, the estimated probability was 21.6% (95% CI=13.4-32.7%)."

For the chip-typed dataset, we plot the cumulative incidence curves,1-S(t) for t between 0 and 100, in Figure 3 (A, C, E and G), where S(t) is the survival function estimated with the Kaplan-Meier method. The results for t=80 are in columns **c** and **d** in the table above.

Comparing the estimates between the two datasets (columns a and c in the table), we observe that the chip-typed dataset has an overrepresentation of cancer cases for breast and prostate cancer and underrepresentation of lung cancer cases, compared to the population level data. To account for the observed bias, the y-axes on the right of Figures 3.A, C, E and G are scaled by the ratio of the cumulative incidence at age 80 for all individuals born in Iceland between 1935-1985 (**a**) to the cumulative incidence at age 80 in the chip-typed dataset (**c**). The ratios are in the last column in the table above.

To clarify, we have now added the table above as Supplementary Table 17, and changed the wording in the Methods and added information to the Figure 3 legend (changes highlighted in yellow):

Probability of cancer diagnosis

The probability of being diagnosed with specific types of cancers in the Icelandic population was computed for all individuals in our Icelandic genealogical database who were born in Iceland between 1935 and 1985 (n=234K). We note that 135K (58%) of them have been chip-typed. For each cancer type, the time variable t was defined as the age at cancer diagnosis or follow-up time in years for those without diagnosis. The probability of being diagnosed before age 80 (t=80) was computed as $P(T \leq t)=1-S(t)$, where S(t) is the survival function. We utilized the Kaplan-Meier estimator with the 'survfit' function in R, from the 'survival' package, to compute S(t). To estimate the corresponding probability of being diagnosed before age 80 for carriers, we applied the estimated OR from the meta-analysis to the population derived estimates.

The cumulative incidence curves, 1-S(t), $t \in [0;100]$, are displayed in Figure 3.A, C, E and G for carriers and non-carriers in the Icelandic chip-typed dataset (N=173K). The chip-typed dataset has an overrepresentation of cancer cases for breast and prostate cancer and an underrepresentation of lung cancer cases, compared to the Icelandic population level data. To account for the observed bias, the y-axes on the right of Figures 3.A, C, E and G are scaled by the ratio of the cumulative incidence at age 80 for all individuals born in Iceland between 1935-1985 to the cumulative incidence at age 80 in the chip-typed dataset (Supplementary Table 17).

Figure 3. The cumulative incidence of cancer among carriers and non-carriers. The cumulative incidence curves, 1-S(t), where S(t) is the survival function estimated with the Kaplan-Meier method, are shown for **A-B;** breast cancer among carriers and non-carriers of LOF+missense variants in *PPP1R15A* in the Icelandic chip-typed dataset and the UK Biobank, **C-D;** prostate cancer among carriers and non-carriers of LOF+missense variants in *BIK* in the Icelandic chip-typed dataset and the UK Biobank, **E-F;** colorectal cancer among carriers and non-carriers of LOF variants in *ATG12* in the Icelandic chip-typed dataset and the UK Biobank and **G-H;** lung cancer among carriers and non-carriers of LOF variants in *CMTR2* in the Icelandic chip-typed dataset and the UK Biobank. In A, C, E, G the cumulative incidence is shown among chip-typed individuals in the Icelandic dataset (N=173,025), and the right y-axis is scaled by the ratio of cumulative incidence at age 80 for all individuals born in Iceland between 1935-1985 (Methods) to the cumulative incidence at age 80 in the chip-typed dataset. The ratios are 0.91 in A, 0.85 in C, 0.98 in E and 1.28 in G (Supplementary Table 17). In B, D, F, H the cumulative incidence is shown among white British/Irish individuals in the UK Biobank (N=431,079).

Decision Letter, third revision:

2nd Aug 2024

Dear Dr Ivarsdottir,

Thank you for submitting your revised manuscript "Gene-based burden test yields rare germline variants in six novel cancer genes" (NG-A63260R2). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy our editorial and formatting guidelines.

If the current version of your manuscript is in a PDF format, please email us a copy of the file in an editable format (Microsoft Word or LaTex)-- we can not proceed with PDFs at this stage.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Safia Danovi, PhD
Senior Editor, Nature Genetics
ORCiD: 0009-0007-7822-5479

Reviewer #3 (Remarks to the Author):

The methods have been clarified and now look consistent with the results and figures. The figures look correct now.

I still think that this is a rather convoluted way of generating cumulative risk graphs (for the Icelandic population). It would be easier just to apply the relative risk estimates for carriers to the population rates (the relative risks are anyway implicitly assumed to be independent of age in the correction) and the authors could consider just doing that. But the difference is minor and the methods are clear.

Final Decision Letter:

30th Sep 2024

Dear Dr Ivarsdottir,

I am delighted to say that your manuscript "Gene-based burden tests of rare germline variants identify six cancer susceptibility genes" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: https://www.nature.com/authors/policies/embargo.html

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact press@nature.com.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that *Nature Genetics* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative

Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to protocols.io. protocols.io is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available and are assigned DOIs for ease of citation. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated workspace to collect all your lab Protocols. By uploading your Protocols to protocols.io, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at https://protocols.io. Further information can be found at https://www.protocols.io/help/publish-articles.

Sincerely,

Safia Danovi, PhD

Senior Editor, Nature Genetics
ORCiD: 0009-0007-7822-5479