

Multi-Ancestry Transcriptome-wide Association Studies Uncover Insights into Breast Cancer Genetics and Biology

Corresponding Author: Professor Wei Zheng

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors present a comprehensive multi-ancestry TWAS framework integrating transcriptomic and genomic data from normal breast tissues across European, Asian, and African ancestries. Ancestry-specific models for gene expression, alternative splicing (AS), and polyadenylation (APA) were applied to GWAS summary statistics from three major breast cancer consortia. The study identified 290 risk-associated genes, including 103 novel TWAS findings and 46 genes located >500Kb from known GWAS loci. Conditional analyses prioritized 152 genes as putative targets for 69 GWAS loci, supported by functional validation using CRISPR-Cas9, siRNA knockdown, and single-cell RNA-Seq data.

The manuscript presents a timely and well-structured contribution to multi-ancestry TWAS methodology in breast cancer research. I appreciate the authors' efforts in integrating diverse molecular phenotypes and ancestries. The following comments are intended to further strengthen the clarity and interpretability of the findings.

Major Comments

1. Although the study includes three major ancestry groups, the data structure remains skewed toward European ancestry. The authors applied inverse-variance weighted meta-analysis to integrate S-PrediXcan results across ancestries and reported ancestry-specific findings for gene expression, AS, and APA. While this is a statistically sound approach, the substantial sample size imbalance—particularly the limited representation of African ancestry—raises concerns about whether the meta-analytic signals are disproportionately driven by European cohorts. To enhance interpretability, I recommend the authors clarify the ancestry contributions to the meta-analysis results and consider discussing or applying multi-ancestry fine-mapping tools such as MA-FOCUS or TESLA. These methods can help distinguish shared causal genes from ancestry-specific signals and mitigate LD-driven biases, thereby improving the biological relevance and generalizability of the findings.
2. The authors employed elastic-net for prediction modeling. While elastic-net has demonstrated strong performance in TWAS (e.g., Gamazon et al., their ref 50), other TWAS frameworks such as FUSION incorporate multiple modeling strategies (e.g., LASSO, BLUP, TOP1) and select the best-performing model per gene/events. It would be helpful if the authors could briefly justify the use of elastic-net exclusively and discuss its limitations relative to alternative approaches.
3. Gene expression levels were predicted using genetic variants within ± 500 Kb of gene. The rationale for selecting this window size is unclear. While 100Kb and 1Mb are commonly used in TWAS studies, the choice of a fixed 500Kb cutoff may be arbitrary and may not reflect the biological reality of long-range regulation. Moreover, the number of modeled genes (~20,000) differs substantially from the number of splicing events (likely much higher) and APA events (likely fewer), which introduces distinct computational burdens using same fixed 500Kb. Furthermore, to improve biological interpretability, the authors may consider discussing alternative SNP-to-gene assignment strategies, such as H-MAGMA (PMID: 32152537).
4. Using all genetic variants within a fixed ± 500 kb window may oversimplify the modeling process. Have the authors considered selecting SNPs based on their association with molecular phenotypes (expression/splicing/APA) and taking their contribution to trait heritability into account (eg. choosing events regulated by snps)? To further contextualize the TWAS results, I suggest estimating SNP-based heritability for each trait using LDSC across ancestries, and considering decomposing or discussing the proportion mediated by gene expression using MESO (Yao et al., Nat Genet 2020). These

analyses could help clarify the extent to which TWAS signals reflect expression-mediated effects and whether such effects vary across tissues or populations.

Minor Comments

1. On line 241, the authors stratified associations by estrogen receptor (ER) status, which is biologically and clinically meaningful. However, for readers less familiar with breast cancer subtypes, it would be helpful if the Introduction briefly explained the rationale for ER-based stratification, including its relevance to disease heterogeneity, treatment response, and molecular mechanisms.

2. Throughout the manuscript, the term “multi-ancestry TWAS” could benefit from clearer definition. It may refer to either (i) single-ancestry TWAS conducted separately across multiple ancestries, or (ii) a unified multi-ancestry TWAS framework that integrates data to produce a single result. Since both approaches appear to be used in this study, especially in summary statements (e.g., line 236: “In total, we identified 263 genes”; line 253: “In total, 290 genes were identified”), I recommend the authors clarify the analytical framework behind each reported result.

3. In main figures and tables, gene names should be italicized in accordance with standard formatting conventions.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

This manuscript by Ping et al presents a large transcriptome-wide association study of breast cancer. The authors have made a significant effort to address the limitations of previous TWAS by including diverse populations and by integrating gene expression, alternative splicing and alternative polyadenylation analyses. Using transcriptomic prediction models with GWAS data, the authors identified 290 candidate susceptibility genes, including more than 100 novel findings and 46 genes outside the known GWAS loci. Functional follow-up using single-cell RNA-seq data and in vitro assays were used to provide support of the biological plausibility of identified genes. The study is interesting and is relevant to the breast cancer genetics field, and the scale and comprehensiveness are commendable. The study has sound statistical methods. I do, however, have a few comments that I believe could improve the manuscript.

The use of SNPs within a ± 500 kb window to predict gene expression is standard in TWAS, but it assumes cis-regulation is the main driver of transcriptomic variation. It would be helpful for the authors to discuss the potential contribution of trans-eQTLs, which can influence expression, splicing and APA (GTEx Consortium, Science (2020)).

The use of Bonferroni correction is appropriate for a discovery study, but it is worth noting that Bonferroni assumes independence of tests. In practice, gene expression is often correlated due to co-regulation, meaning Bonferroni can be overly conservative. The authors may wish to acknowledge this as a minor limitation.

The sample size imbalance (more European than Asian and African) is acknowledged, but could the authors quantify the relative power across these groups and discuss on how this could affect the likelihood of discovering ancestry-specific signals?

Conditioning only on the nearest lead GWAS SNP (although the study used the stronger association) may inadequately control for multiple independent signals within loci. The authors may comment this in the discussion section.

In the cell-type specificity analysis of scRNA-seq data, were any genes specifically associated with risk in the ER-positive analysis also enriched in luminal cells, given their proposed role in luminal cancers? This could be an interesting point for the discussion section.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

Summary:

The study focuses on building gene prediction models by integrating genomic and transcriptomic data (TWAS) from normal breast tissue across ancestries (African, Asian, European) with the aim to study gene-trait associations. The study also highlights the significance of posttranscriptional gene regulation programs like alternative splicing (AS) and alternative polyadenylation (APA) and predicts their role in influencing gene expression and phenotypes. Authors argue that due to the limited availability of transcriptomic data from normal breast tissue, previous studies built their gene prediction models by utilizing transcriptomic data from other tissues. Therefore, to build more accurate and ancestry-specific models to predict gene expression, AS, and APA.

Major comments:

I commend the authors on their inclusion of non-European ancestry participants. Presumably this is to not only identify if the genes associated with breast cancer risk in European-ancestry participants are replicated in Asian and African-ancestry participants, but to identify if Asian and African-ancestry is associated with different mechanisms of carcinogenesis that lead to poor outcomes. It would add value if the authors could expand on the novel findings identified from analysis of Asian and African-ancestry participants and how confident can the authors be in these findings based on the sample size of these groups of participants?

For evaluating the validity of gene expression prediction models authors compared the predicted gene expression with the observed gene expression. In Figures S2A and S2B both graphs seem identical; however, R2 of 0.87 and 0.81 respectively are calculated.

Were the genes that exhibited association with breast cancer risk based on ER status, also predicted to undergo AS or APA? Which cell signaling pathways the predicted genes associate with? This should be discussed.

AS and APA greatly influence 3'UTR length which in turn determines mRNA fate by affecting mRNA stability, translation, and localization. While the study lists the breast cancer risk genes (115 genes) that are predicted to undergo AS and APA, whether or how these potential post-transcriptional modifications influence overall expression of these genes are not discussed. Moreover, no experimental evidence/validation is provided to suggest that mRNAs harbouring these modifications contribute to proliferative phenotype or the opposite.

Regarding the siRNA knockdown experiments:

- Gene knockdown results in Table S9 lack an explanation for how to interpret these scores. Which of these results represent a significant decrease in viability compared to control? How many different siRNA were used to target each gene (a table should be provided with this information)? By using multiple siRNAs targeting the same gene this could further support the data and negate the artifacts caused by siRNA off targeting. How was knockdown validated for each siRNA and each gene? How was off targeting measured or controlled? How many replicates were conducted per siRNA? What is the variability in viability scores across replicates?
- The authors use 3 breast cancer cell lines representing different subtypes of breast cancer. It would add value if they expanded on the differences in genetic vulnerabilities identified between subtypes and how this is consistent with targeted therapies used to treat the different subtypes.
- The siRNA knockdown experiments were performed on only a limited number of cell lines, and it is unclear how consistent these results would be by breast cancer subtype if additional cell lines representing these 3 subtypes were added. Furthermore, it is unclear how these results would translate to more representative models of breast cancer such as patient-derived tumor xenografts or with in vivo experiments.

Regarding the scRNAseq differential expression analysis:

- The authors should present a UMAP or other similar visualization of the dataset.
- How were the different cell types annotated? The authors should include a table of the marker genes used for cell type annotation.
- The authors group all luminal epithelial cells into a single cell type but it is well known there are luminal progenitor cells and mature luminal cells. They can be distinguished in scRNAseq data, demonstrate distinct functional properties, with luminal progenitor cells thought to be the cell of origin for certain types of breast cancer. BRCA1 carriers have also been found to have a perturbed luminal progenitor cell fraction. In their analysis, the authors should separate luminal epithelial cells into mature luminal and luminal progenitor cells, and also provide information on whether any of the normal samples include high risk mutations. If they do, analysis of how the differential expression analysis differs between groups of patients with different high risk mutations should be included.
- Results are summarized in Table S10 but the authors miss an opportunity to elaborate on what new or meaningful information can be concluded from this analysis.

The authors state in the discussion: "Our study provided substantial novel information on the potential mechanisms underlying breast cancer risk and significantly expanded the catalog of candidate breast cancer susceptibility genes." I agree they do expand the catalog of candidate breast cancer susceptibility genes, but they do not elaborate on how this provides new or deeper insight into breast cancer risk. The authors also include analysis of AS and APA across the transcriptome, but how has this yielded new understanding of breast cancer risk? How has this study improved our understanding of breast cancer in a meaningful way?

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have adequately addressed all the concerns raised in my previous review. I appreciate their efforts in carefully responding to my comments and revising the manuscript. At this stage, I only have a few minor formatting concerns:

(1) Gene names should be italicized throughout the manuscript, including in the supplementary tables. While this is not a major issue, following the standard conventions for gene nomenclature would further enhance readability.

(2) In Tables 2 and 3, the formatting of the P-value column is inconsistent: some small values are presented in scientific notation (e.g., 7.60×10^{-12} , 6.04×10^{-54}), while others are shown as decimals (e.g., 0.54, 0.19, 0.30) instead of in scientific notation (e.g., 5.40×10^{-1} , 1.90×10^{-1} , 3.00×10^{-1}). Although this does not affect the scientific content, it may improve clarity and readability to adopt a consistent format—for example, presenting all P-values in scientific notation. This is a minor suggestion and can be considered at the authors' discretion.

(3) In the Methods section (PDF line 668), the 'P' in the notation "with a P-value cutoff of 0.05" should be italicized to remain consistent with the formatting used throughout the manuscript. Ensuring uniform formatting of statistical symbols would improve clarity and professionalism.

Overall, I find the manuscript substantially improved and suitable for publication after these minor adjustments. Addressing these formatting details would further enhance the quality of the final version.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

In this second round of revision, the authors have satisfactorily addressed all of my previous comments. I have no further points to raise.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have satisfactorily addressed my concerns in their revised manuscript.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewer Comments

Reviewer #1 (Remarks to the Author):

The authors present a comprehensive multi-ancestry TWAS framework integrating transcriptomic and genomic data from normal breast tissues across European, Asian, and African ancestries. Ancestry-specific models for gene expression, alternative splicing (AS), and polyadenylation (APA) were applied to GWAS summary statistics from three major breast cancer consortia. The study identified 290 risk-associated genes, including 103 novel TWAS findings and 46 genes located >500Kb from known GWAS loci. Conditional analyses prioritized 152 genes as putative targets for 69 GWAS loci, supported by functional validation using CRISPR-Cas9, siRNA knockdown, and single-cell RNA-Seq data.

The manuscript presents a timely and well-structured contribution to multi-ancestry TWAS methodology in breast cancer research. I appreciate the authors' efforts in integrating diverse molecular phenotypes and ancestries. The following comments are intended to further strengthen the clarity and interpretability of the findings.

Response: We thank the reviewer for the positive assessment and for the insightful comments, which have significantly enhanced the quality and interpretability of our study. Below, we provide a point-by-point response to each comment.

Major Comments

1. Although the study includes three major ancestry groups, the data structure remains skewed toward European ancestry. The authors applied inverse-variance weighted meta-analysis to integrate S-PrediXcan results across ancestries and reported ancestry-specific findings for gene expression, AS, and APA. While this is a statistically sound approach, the substantial sample size imbalance—particularly the limited representation of African ancestry—raises concerns about whether the meta-analytic signals are disproportionately driven by European cohorts. To enhance interpretability, I recommend the authors clarify the ancestry contributions to the meta-analysis results and consider discussing or applying multi-ancestry fine-mapping tools such as MA-FOCUS or TESLA. These methods can help distinguish shared causal genes from ancestry-

specific signals and mitigate LD-driven biases, thereby improving the biological relevance and generalizability of the findings.

Response: We thank the reviewer for the helpful feedback. We agree that the meta-analytic signals could be disproportionately driven by the European-ancestry data given the large sample size of this ancestral group. To address this, we present the heterogeneity statistics (I^2 and P_{Het}) in Supplementary Tables, showing that no apparent heterogeneity is present in most signals. We appreciate the suggestion to use multi-ancestry fine-mapping methods like MA-FOCUS or TESLA. In fact, we had devoted substantial efforts to explore these approaches to attempt to distinguish shared from ancestry-specific signals. Unfortunately, we found that these methods were not robustly applicable to our data. These tools generally require adequately powered summary statistics from each participating ancestry to reliably deconvolve LD and fine-map signals. However, due to a small sample size of our non-European groups, the models either failed to converge or produced unstable results. Furthermore, there were considerable differences in prediction model accuracy across ancestral groups for many genes, and it was difficult to compare association signals across these groups. We concluded that our data did not meet the requirements necessary for these methods to yield reliable biological insights. As a conservative alternative, we conducted single-ancestry fine-mapping with FOCUS using ancestry-specific LD (EUR/ASN/AFR), reported in **Table S14** and added the results to the Results section (page 11, line 1-12). We also explain the rationale and limitations in Methods (page 26, line 8-18) and Discussion (page 16, line 20-21) sections.

2. The authors employed elastic-net for prediction modeling. While elastic-net has demonstrated strong performance in TWAS (e.g., Gamazon et al., their ref 50), other TWAS frameworks such as FUSION incorporate multiple modeling strategies (e.g., LASSO, BLUP, TOP1) and select the best-performing model per gene/events. It would be helpful if the authors could briefly justify the use of elastic-net exclusively and discuss its limitations relative to alternative approaches.

Response: We selected the elastic-net model because it balances sparsity and polygenicity and performed well in cross-validation across gene expression, splicing, and APA models. Elastic-net is particularly well-suited for building genetic prediction models of gene expression because it

effectively balances the sparse nature of cis-eQTLs (where only a subset of local SNPs is causal) with the highly correlated structure of genetic data. While we acknowledge that multi-model frameworks like FUSION can sometimes improve predictive performance for specific genes by testing a suite of models (e.g., LASSO), we decided to use a consistent, robust, and computationally efficient framework applied uniformly across all genes, different transcriptomic events, and ancestries. This consistency simplifies interpretation and reduces the potential for overfitting or "model-shopping." Importantly, we rigorously validated the performance of our models. As shown in **Figure S2** and described in the manuscript, a cross-dataset validation between models trained on Komen data and tested on GTEx data yielded a high correlation coefficient of 0.87, confirming the robustness of our chosen approach. Furthermore, within our framework, when multiple models were available for a given transcriptomic event, we selected the one with the highest prediction R^2 , ensuring we used the best-performing model within our consistent methodology. We have now added a sentence to the Methods section to briefly clarify the rationale for using elastic-net (page 24, line 24-25). We also have deposited our data to dbGaP (phs003535) so that other groups can use various model building methods in their future studies.

3. Gene expression levels were predicted using genetic variants within $\pm 500\text{Kb}$ of gene. The rationale for selecting this window size is unclear. While 100Kb and 1Mb are commonly used in TWAS studies, the choice of a fixed 500Kb cutoff may be arbitrary and may not reflect the biological reality of long-range regulation. Moreover, the number of modeled genes ($\sim 20,000$) differs substantially from the number of splicing events (likely much higher) and APA events (likely fewer), which introduces distinct computational burdens using same fixed 500Kb. Furthermore, to improve biological interpretability, the authors may consider discussing alternative SNP-to-gene assignment strategies, such as H-MAGMA (PMID: 32152537).

Response: We appreciate the reviewer's concern. We decided to use $\pm 500\text{kb}$ window as it balances the capture of local cis-regulatory effects and computational efficiency. While long-range regulation exists, the vast majority of cis-eQTLs are within this window, and thus expanding the window to 1Mb would increase unnecessary computation burden. Narrowing the window to 100kb would miss some cis-eQTLs.

The reviewer's suggestion to consider H-MAGMA is appreciated. H-MAGMA leverages 3D chromatin interaction data (Hi-C) to improve gene-based analyses by assigning non-coding GWAS SNPs to their cognate genes, which may be distal, which is beyond our current study's scope. We've acknowledged the potential of integrating Hi-C data for model building for future work in the Discussion section (page 17, line 12-15). Again, our data are shared with the scientific community for other groups to explore various ways of model building.

4. Using all genetic variants within a fixed $\pm 500\text{kb}$ window may oversimplify the modeling process. Have the authors considered selecting SNPs based on their association with molecular phenotypes (expression/splicing/APA) and taking their contribution to trait heritability into account (eg. choosing events regulated by snps)? To further contextualize the TWAS results, I suggest estimating SNP-based heritability for each trait using LDSC across ancestries, and considering decomposing or discussing the proportion mediated by gene expression using MESOC (Yao et al., Nat Genet 2020). These analyses could help clarify the extent to which TWAS signals reflect expression-mediated effects and whether such effects vary across tissues or populations.

Response: We appreciate the reviewer's excellent suggestion. As detailed in our Methods, we employed the elastic-net model, which inherently performs variable selection by shrinking the coefficients of non-predictive variants to zero while retaining correlated variants that contribute to the phenotype. Therefore, although we started with all variants in a $\pm 500\text{kb}$ window to capture local cis-architecture, the final models only utilized a subset of SNPs specifically associated with the molecular phenotypes. Estimating breast cancer heritability that is mediated by the genetically regulated transcriptome is a relevant scientific question. However, performing a comprehensive, multi-ancestry heritability partitioning analysis using methods like LDSC and MESOC is a substantial research project requiring different analytical pipelines and careful interpretation of cross-ancestry genetic correlations. While important, it extends beyond our manuscript's focus, particularly given the substantial amount of results included in this manuscript. We've added a sentence to our Discussion highlighting it as a valuable next step for future research.

Minor Comments

1. *On line 241, the authors stratified associations by estrogen receptor (ER) status, which is biologically and clinically meaningful. However, for readers less familiar with breast cancer subtypes, it would be helpful if the Introduction briefly explained the rationale for ER-based stratification, including its relevance to disease heterogeneity, treatment response, and molecular mechanisms.*

Response: We agree and have added a few sentences to the Introduction (page 5, line 21 - page 6, line 8) to explain the biological and clinical importance of stratifying breast cancer by ER status.

2. *Throughout the manuscript, the term “multi-ancestry TWAS” could benefit from clearer definition. It may refer to either (i) single-ancestry TWAS conducted separately across multiple ancestries, or (ii) a unified multi-ancestry TWAS framework that integrates data to produce a single result. Since both approaches appear to be used in this study, especially in summary statements (e.g., line 236: “In total, we identified 263 genes”; line 253: “In total, 290 genes were identified”), I recommend the authors clarify the analytical framework behind each reported result.*

Response: We thank the reviewer for pointing out this ambiguity. Our study indeed uses a hybrid approach. We have revised the Methods section to explicitly define our two-stage framework: 1) conducting ancestry-specific TWAS using ancestry-specific prediction models and GWAS data, and 2) combining these results via an inverse-variance-weighted meta-analysis to identify robust, cross-ancestry signals, while also reporting significant findings from the individual ancestry analyses. Therefore, multi-ancestry TWAS in our study refers the use of multi-ancestry genomic and transcriptomic data for TWAS, including analyses conducted using data either from a single ancestry or more than one ancestral group (page 22, line 22-27).

3. *In main figures and tables, gene names should be italicized in accordance with standard formatting conventions.*

Response: We appreciate the reviewer's attention to this detail. We have performed a thorough review of the entire manuscript, including all figures, main text, and supplementary tables, and have ensured that all gene names are now correctly italicized.

Reviewer #2 (Remarks to the Author):

This manuscript by Ping et al presents a large transcriptome-wide association study of breast cancer. The authors have made a significant effort to address the limitations of previous TWAS by including diverse populations and by integrating gene expression, alternative splicing and alternative polyadenylation analyses.

Using transcriptomic prediction models with GWAS data, the authors identified 290 candidate susceptibility genes, including more than 100 novel findings and 46 genes outside the known GWAS loci. Functional follow-up using single-cell RNA-seq data and in vitro assays were used to provide support of the biological plausibility of identified genes. The study is interesting and is relevant to the breast cancer genetics field, and the scale and comprehensiveness are commendable. The study has sound statistical methods. I do, however, have a few comments that I believe could improve the manuscript.

Response: We are grateful to reviewer for the positive assessment and constructive comments, which have helped us to improve the manuscript. We have addressed each point below.

The use of SNPs within a ± 500 kb window to predict gene expression is standard in TWAS, but it assumes cis-regulation is the main driver of transcriptomic variation. It would be helpful for the authors to discuss the potential contribution of trans-eQTLs, which can influence expression, splicing and APA (GTEx Consortium, Science (2020)).

Response: We appreciate the reviewer's point. Our study focused on cis-eQTLs, a common approach in TWAS model building that prioritizes robustly modeled local effects. We agree that trans-eQTLs could contribute to breast cancer susceptibility. However, incorporating them into prediction models presents challenges, including an increased multiple testing burden, the need

for large reference panels, and difficulties distinguishing true trans-effects from artifacts. Thus, our study focused on the well-established cis-regulatory architecture. We have added a sentence to the Discussion (page 17, line 7-9) about incorporating trans-acting regulation in future research.

The use of Bonferroni correction is appropriate for a discovery study, but it is worth noting that Bonferroni assumes independence of tests. In practice, gene expression is often correlated due to co-regulation, meaning Bonferroni can be overly conservative. The authors may wish to acknowledge this as a minor limitation.

Response: We appreciate the reviewer's comment. We agree that the Bonferroni correction is conservative. We applied this stringent threshold to minimize false positives. We've added a sentence to the Discussion acknowledging that this method is too conservative and it may miss some true associations.

The sample size imbalance (more European than Asian and African) is acknowledged, but could the authors quantify the relative power across these groups and discuss on how this could affect the likelihood of discovering ancestry-specific signals?

Response: We appreciate this comment. The manuscript's Discussion acknowledges the sample size disparity as a key limitation. While a formal power calculation is complex and also determined by unknown parameters, and thus it varies by genes, such as prediction accuracy of models. Because of a relatively small sample size in non-European populations, some true associations in Asian or African populations may be missed, and a failure to replicate a European ancestry finding in the other groups could be due to limited power rather than a true biological difference. We've expanded on this point in the revised Discussion section.

Conditioning only on the nearest lead GWAS SNP (although the study used the stronger association) may inadequately control for multiple independent signals within loci. The authors may comment this in the discussion section.

Response: Our conditional analysis on the nearest GWAS-identified risk variant was conducted to determine if our TWAS signals were independent of the main known GWAS association. We understand that this approach may not fully resolve associations in complex loci with multiple independent causal variants. Our new fine-mapping analysis (in response to Reviewer #1) addresses this by statistically dissecting these loci to identify likely causal genes, a more sophisticated approach to resolving multiple signals. We clarified the limitation of the single-SNP conditional analysis and noted the utility of advanced methods in the Discussion.

In the cell-type specificity analysis of scRNA-seq data, were any genes specifically associated with risk in the ER-positive analysis also enriched in luminal cells, given their proposed role in luminal cancers? This could be an interesting point for the discussion section.

Response: We thank the reviewer for this insightful suggestion. We cross-referenced our ER-positive TWAS associations with single-cell RNA-seq differential expression results. Several genes significantly associated with ER-positive breast cancer risk were indeed enriched in luminal epithelial cells compared to other cell types (Supplementary Table S10). This further supports the biological relevance of our findings, linking genetic risk for a tumor subtype to its cell of origin. We added a new paragraph to the Discussion to highlight this.

Reviewer #3 (Remarks to the Author):

Summary:

The study focuses on building gene prediction models by integrating genomic and transcriptomic data (TWAS) from normal breast tissue across ancestries (African, Asian, European) with the aim to study gene-trait associations. The study also highlights the significance of posttranscriptional gene regulation programs like alternative splicing (AS) and alternative polyadenylation (APA) and predicts their role in influencing gene expression and phenotypes. Authors argue that due to the limited availability of transcriptomic data from normal breast tissue, previous studies built their gene prediction models by utilizing transcriptomic data from

other tissues, Therefore, to build more accurate and ancestry-specific models to predict gene expression, AS, and APA.

Response: We thank the reviewer for the positive evaluation of our study's aims and design. We have provided detailed responses below.

Major comments:

I commend the authors on their inclusion of non-European ancestry participants. Presumably this is to not only identify if the genes associated with breast cancer risk in European-ancestry participants are replicated in Asian and African-ancestry participants, but to identify if Asian and African-ancestry is associated with different mechanisms of carcinogenesis that lead to poor outcomes. It would add value if the authors could expand on the novel findings identified from analysis of Asian and African-ancestry participants and how confident can the authors be in these findings based on the sample size of these groups of participants?

Response: We agree with the reviewer about the unique contribution of this study. As stated in the manuscript, 46 genes were associated with risk across all three populations, and 70 were identified in two populations. This cross-ancestry replication, despite differences in genetic architectures and model parameters, provides strong evidence that the identified associations are likely to be causal. Also, we also identified 80 associated genes in the ancestry-specific analysis (Table S2-S4).

For evaluating the validity of gene expression prediction models authors compared the predicted gene expression with the observed gene expression. In Figures S2A and S2B both graphs seem identical; however, R2 of 0.87 and 0.81 respectively are calculated.

Response: We appreciate the reviewer's identification of this error. The reviewer is correct that Figures S2A and S2B were inadvertently duplicated in the original submission. This was a clerical error, and we have corrected it.

Were the genes that exhibited association with breast cancer risk based on ER status, also

predicted to undergo AS or APA? Which cell signaling pathways the predicted genes associate with? This should be discussed.

Response: We found association signals in ER-stratified analysis of predicted alternative splicing (AS) or alternative polyadenylation (APA). For instance, our spTWAS linked AS in *NSUN4* and *PEX14* to ER-positive and ER-negative breast cancer risk, respectively (Table S6). Our APA-WAS similarly linked APA in *NSUN4* to ER-positive risk (Table S7). Pathway analyses showed that genes identified in the ER-negative analysis were enriched in the p53 signaling and apoptosis pathways (Table S8). This suggests that post-transcriptional modifications in specific genes may influence these critical cancer pathways to modulate risk for specific breast cancer subtypes. We've added a new paragraph to the Discussion to address this important intersection.

AS and APA greatly influence 3'UTR length which in turn determines mRNA fate by affecting mRNA stability, translation, and localization. While the study lists the breast cancer risk genes (115 genes) that are predicted to undergo AS and APA, whether or how these potential post-transcriptional modifications influence overall expression of these genes are not discussed. Moreover, no experimental evidence/validation is provided to suggest that mRNAs harbouring these modifications contribute to proliferative phenotype or the opposite.

Response: Our study identified 115 genes exclusively through spTWAS or APA-WAS, highlighting the importance of post-transcriptional regulation. We have now expanded our Discussion to elaborate on the well-established biological mechanisms through which these events can act (e.g., altering protein isoforms or mRNA stability). While our siRNA experiments demonstrate that overall knockdown of many of these genes impacts cell proliferation, we acknowledge that these experiments do not distinguish the effects of specific isoforms. We have clarified this limitation and highlighted the need for future functional studies.

Regarding the siRNA knockdown experiments:

- Gene knockdown results in Table S9 lack an explanation for how to interpret these scores. Which of these results represent a significant decrease in viability compared to control? How

many different siRNA were used to target each gene (a table should be provided with this information)? By using multiple siRNAs targeting the same gene this could further support the data and negate the artifacts caused by siRNA off targeting. How was knockdown validated for each siRNA and each gene? How was off targeting measured or controlled? How many replicates were conducted per siRNA? What is the variability in viability scores across replicates?

Response: We have added a new paragraph in the Methods section to the manuscript with a detailed description of the siRNA screening protocol. To summarize here:

- Interpretation: The values in Table S9 are P-values from a t-test comparing the viability of cells treated with the target-specific siRNA to cells treated with a scrambled negative control siRNA. As stated in the manuscript, a result was considered significant if $P < 0.05$.
- Reagents and Replicates: This was a high-throughput screen using a validated library of siRNA pools, with each pool containing multiple siRNAs targeting a single gene to increase knockdown efficiency and reduce off-target effects. All experiments were conducted in triplicate.
- Validation and Controls: For a screen of this scale, direct validation of knockdown efficiency (e.g., by qPCR) for all 216 genes was not feasible. We relied on the use of a pre-validated siRNA library and included appropriate positive (e.g., targeting a known essential gene) and negative (scrambled sequence) controls in every plate to ensure assay performance. The use of siRNA pools is a standard strategy to mitigate off-target effects from any single siRNA sequence. We have clarified these points in the new Supplementary Methods and in the Discussion.

- The authors use 3 breast cancer cell lines representing different subtypes of breast cancer. It would add value if they expanded on the differences in genetic vulnerabilities identified between subtypes and how this is consistent with targeted therapies used to treat the different subtypes.

Response: We used cell lines representing three major subtypes: MCF-7 (ER), SK-BR-3 (HER2), and MDA-MB-231 (TNBC). Upon re-examining our data (Table S9), we found subtype-specific vulnerability. For instance, knockdown of NSUN4, associated with ER-positive

risk, significantly reduced proliferation in the ER-positive MCF-7 cell line ($P=0.01$) but not in the other two. This strengthens the link between NSUN4 and ER-positive cancer biology. We added a paragraph to the Discussion to highlight this and other examples, connecting our genetic findings to subtype-specific cellular dependencies. We also added one more column in the Table S9 to indicate the significant cell-type.

- The siRNA knockdown experiments were performed on only a limited number of cell lines, and it is unclear how consistent these results would be by breast cancer subtype if additional cell lines representing these 3 subtypes were added. Furthermore, it is unclear how these results would translate to more representative models of breast cancer such as patient-derived tumor xenografts or with in vivo experiments.

Response: The siRNA experiments were intended as a first-pass functional screen, and we acknowledge the limitations of using a small number of established cell lines. These models cannot capture the full inter-patient heterogeneity or the complex tumor microenvironment. We have now explicitly stated in the Discussion that these findings require validation in a broader panel of cell lines and, more importantly, in more clinically relevant models such as patient-derived xenografts or organoids.

Regarding the scRNAseq differential expression analysis:

- The authors should present a UMAP or other similar visualization of the dataset.
- How were the different cell types annotated? The authors should include a table of the marker genes used for cell type annotation.

Response: Our analysis used the publicly available and well-annotated single-cell RNA-seq dataset for normal breast tissue from the GTEx project. To ensure robustness and reproducibility, we directly used the expert-curated cell type annotations from the GTEx consortium based on the expression of canonical marker genes. For example, luminal epithelial cells were identified using markers like KRT8 and KRT18, myoepithelial cells using markers like KRT14 and ACTA2, and fibroblasts using markers like DCN and LUM. Detailed visualizations of cell clustering, such as UMAP plots, and the complete list of marker genes are available in the primary GTEx

consortium publication describing this dataset, which we now cite in the Methods section.

Relying on these established, public annotations strengthens the validity of our cell-type-specific findings.

- The authors group all luminal epithelial cells into a single cell type but it is well known there are luminal progenitor cells and mature luminal cells. They can be distinguished in scRNAseq data, demonstrate distinct functional properties, with luminal progenitor cells thought to be the cell of origin for certain types of breast cancer. BRCA1 carriers have also been found to have a perturbed luminal progenitor cell fraction. In their analysis, the authors should separate luminal epithelial cells into mature luminal and luminal progenitor cells, and also provide information on whether any of the normal samples include high risk mutations. If they do, analysis of how the differential expression analysis differs between groups of patients with different high risk mutations should be included.

Response: We appreciate that distinguishing luminal progenitor from mature cells is helpful, but this study used validated cell type annotations from GTEx for robustness. A detailed sub-clustering of the luminal compartment was beyond our scope. Information on high-risk germline mutations for GTEx donors was unavailable, so we couldn't analyze it. We have added this as a limitation in our Discussion (page 17, line 15-18).

- Results are summarized in Table S10 but the authors miss an opportunity to elaborate on what new or meaningful information can be concluded from this analysis.

Response: We have expanded the Discussion to clarify the significance of this analysis. Specifically, we emphasize that these results establish a crucial mechanistic link between a statistical genetic association and the cell type from which a cancer subtype originates. For instance, finding that a gene conferring risk for ER-positive cancer is most highly expressed in normal luminal cells provides strong biological evidence that its dysregulation within this cell lineage is a key event in tumorigenesis.

The authors state in the discussion: "Our study provided substantial novel information on the

potential mechanisms underlying breast cancer risk and significantly expanded the catalog of candidate breast cancer susceptibility genes.” I agree they do expand the catalog of candidate breast cancer susceptibility genes, but they do not elaborate on how this provides new or deeper insight into breast cancer risk. The authors also include analysis of AS and APA across the transcriptome, but how has this yielded new understanding of breast cancer risk? How has this study improved our understanding of breast cancer in a meaningful way?

Response: We’ve revised the Discussion to highlight our novel contributions: 1) We identified 115 risk-associated genes exclusively through alternative splicing and APA analysis, which conventional TWAS missed. Post-transcriptional regulation is a critical mechanism mediating genetic risk. 2) Our multi-ancestry approach provides stronger evidence for causality and uncovers ancestry-specific signals, which has implications for equitable risk prediction. 3) We prioritize the most likely causal gene(s) within an associated genomic region using fine-mapping, a critical step toward testable biological hypotheses. These points are now more prominent in the Discussion.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We thank the reviewer for the contribution to the peer-review process. We appreciate the time and effort in evaluating our manuscript and support the initiative to provide training and recognition for early-career researchers in this critical scientific activity.

Response to Reviewer Comments

Reviewer #1 (Remarks to the Author):

The authors have adequately addressed all the concerns raised in my previous review. I appreciate their efforts in carefully responding to my comments and revising the manuscript. At this stage, I only have a few minor formatting concerns:

Response: We thank the reviewer for the positive assessment and for the insightful comments, which have significantly enhanced the quality and interpretability of our study. Below, we provide a point-by-point response to each comment.

(1) Gene names should be italicized throughout the manuscript, including in the supplementary tables. While this is not a major issue, following the standard conventions for gene nomenclature would further enhance readability.

Response: We thank the reviewer for this helpful suggestion. We have carefully reviewed the manuscript and supplementary materials and have italicized human gene symbols throughout the main text, tables, figure legends, and supplementary tables/data files, following standard gene nomenclature conventions. We also checked the formatting of gene symbols in Tables 1-3 and in all relevant supplementary tables to ensure consistency.

(2) In Tables 2 and 3, the formatting of the P-value column is inconsistent: some small values are presented in scientific notation (e.g., 7.60×10^{-12} , 6.04×10^{-54}), while others are shown as decimals (e.g., 0.54, 0.19, 0.30) instead of in scientific notation (e.g., 5.40×10^{-1} , 1.90×10^{-1} , 3.00×10^{-1}). Although this does not affect the scientific content, it may improve clarity and readability to adopt a consistent format—for example, presenting all P-values in scientific notation. This is a minor suggestion and can be considered at the authors' discretion.

Response: We thank the reviewer for noting this formatting inconsistency. We agree that using a consistent format improves readability. We have revised Tables 2 and 3 so that all P values are

presented consistently in scientific notation. We also reviewed the corresponding supplementary tables/data files and standardized P-value formatting where applicable.

(3) In the Methods section (PDF line 668), the 'P' in the notation "with a P-value cutoff of 0.05" should be italicized to remain consistent with the formatting used throughout the manuscript. Ensuring uniform formatting of statistical symbols would improve clarity and professionalism.

Response: We thank the reviewer for this careful reading. We have revised the Methods section so that the statistical symbol P is italicized in the phrase "with a P-value cutoff of 0.05." We also reviewed the manuscript for similar instances and standardized the formatting of statistical symbols where needed.

Overall, I find the manuscript substantially improved and suitable for publication after these minor adjustments. Addressing these formatting details would further enhance the quality of the final version.

Reviewer #2 (Remarks to the Author):

In this second round of revision, the authors have satisfactorily addressed all of my previous comments. I have no further points to raise.

Reviewer #3 (Remarks to the Author):

The authors have satisfactorily addressed my concerns in their revised manuscript.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.