nature portfolio

Peer Review File

A Cross-population Compendium of Gene–Environment Interactions

Corresponding Author: Professor Yukinori Okada

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
A Cross-population Compendium of Gene–Environment Interactions
Shinichi Namba et al.

This study explores gene–environment (G×E) interactions using large biobank data from 543,732 European and Japanese individuals, covering 47 phenotypes and 12 environmental factors. It provides a global overview of G×E interactions, focusing on replication, cross-population consistency, and the pleiotropic effects of genes on disease risks. The study highlights the role of G×E interactions in understanding phenotypic plasticity, identifying how age, sex, and lifestyle factors contribute to specific loci, including a notable example of reverse-causal G×E interaction related to disease-driven dietary changes. Through genome-wide interaction studies, the authors address the challenge of missing heritability within the G×E framework. They also illustrate the influence of G×E interactions on polygenic prediction accuracy and cross-population applicability. The authors contend that this study serves as a vital resource for understanding the dynamics of genetic associations, with significant implications for complex trait biology, personalized medicine, and drug development.

This is an important and well-written study that provides a comprehensive analysis. While the manuscript is strong overall, I have a few comments that may be critical to address. Incorporating these suggestions could significantly enhance the clarity and impact of the current version.

1. Causal Directions: It remains unclear how the authors establish the causal direction in their analysis. For instance, could the genetic effects on alcohol consumption (such as rs672's impact on drinking behavior) be influenced by factors like AST or Hb levels? The fact that drinking behavior was not measured prior to assessing AST or Hb levels presents a potential limitation in the study design. Or, please clarify if this is not the case. This limitation arises from the inability to establish clear temporal causality—specifically, whether drinking behavior influenced AST and Hb levels or whether pre-existing conditions or genetic predispositions (such as rs671) may have influenced both alcohol consumption and the biomarkers. To resolve this, the authors might consider conducting a Mendelian Randomization (MR) study to establish a causal relationship and avoid ambiguity around the directionality of effects.

2. Reverse Causality: The authors mention, "In summary, we identified the G×E interaction caused by a reverse causal relationship from the disease to the environment" (line 393). However, while the study suggests reverse causality—such as atrial fibrillation (AF) leading to changes in natto consumption due to warfarin treatment—the evidence is suggestive rather than conclusive. Several alternative mechanisms, such as confounding factors, pleiotropy, incomplete data on treatment timing, or non-dietary factors, could also explain the observed G×E interaction. To strengthen their findings and rule out alternative explanations, the authors could consider applying a causal inference analysis, such as MR or other techniques, to clarify the causal pathways and ensure robust conclusions.

3. There is a risk of generating spurious signals in G×E studies when adjusting for covariates, particularly due to collider bias. I notice that multiple covariates were used in the analysis. How do the authors ensure that the covariates selected do not introduce spurious signals? A detailed explanation of the covariate selection process, particularly how potential colliders

were handled, would be beneficial.

4. In relation to #3, I suggest validating the estimated G×E effects by projecting them from the discovery dataset to an independent target dataset. This would help assess whether the estimated G×E interactions have significant predictive value in an external dataset. The use of G×Eprs method, as outlined in https://github.com/DoviniJ/GxEprs (Genetic Epidemiology, 48(2), 85-100), could be a useful approach for this purpose.

5. The prediction performance of polygenic scores (PGS) is well demonstrated. However, the analyses presented focus more on performance under homogeneous environmental conditions, rather than explicitly modeling G×E interactions. Could the authors conduct similar prediction analyses comparing models with and without G×E interactions? Additionally, do any of the PGS rankings change when incorporating G×E modeling? This comparison would provide valuable insights into the impact of G×E interactions on prediction accuracy.

6. The omic-wide interaction models are not thoroughly described in the Methods section. The authors should provide more detailed information about the model, including its implementation and underlying assumptions. This is particularly important because non-genomic omics data are more susceptible to being influenced by outcomes compared to genomic variables. A precise and comprehensive description of the model would help address potential concerns related to reverse causality and ensure clarity for reproducibility.

Referee #2

(Remarks to the Author)
To the author

This article by Namba et. al. demonstrates an excellent application of detecting gene environment interactions in two cohorts. It expands on current knowledge by providing an estimate of cross population replication, estimating missing heritability attributed to GxE, and taking a closer examination on the biology behind top GxEs. The authors also utilize single-cell transcriptomics, metabolomics, and proteomics data to connect interactions to molecular markers. Strengths include methodology, clear writing, and beautiful visualizations. The analyses and results are not as comprehensive or novel as claimed, and this is my largest concern for publication in Nature.

Comments

1. The authors greatly over-state the impact of their findings with descriptors such as “compendium”, “atlas”, “[phenome-, environment-, etc] -wide”, “comprehensive”, “catalog”, and “cross-population”. They only investigated 47 phenotypes, 12 environments (which also include less “environmental” exposures such as age and sex), and 2 biobanks of 2 different ancestries, making a much more focused analysis, and do not provide a “catalog” or “atlas” resource for readers to interrogate a “comprehensive” set of GxEs.
2. The approach for identifying significant GxE is confusing and unclear. Lines 217 – 225, and lines 911- 940: it appears that an E was considered significant it if it improved the overall significant model, but not based on whether the interaction itself was significant. Please clarify whether Es were also examined independently before being considered significant. There is a big difference in interpretation if an individual interaction term is significant versus the overall model fit is improved due to the inclusion of an interaction term.
3. The study-wide significance threshold does not account for the number of environmental interactors in addition to the number of traits. The appropriate correction is something closer to 5x10-8/(47 phenotypes * 12 environments) = 8.8x10-11. You could also argue 2 discovery cohorts and an additional joint environment model could make this even more stringent. Although hard to tell from ST5, this takes significant loci down to something closer to 18. These strict multiple testing burdens are why few GxE have been discovered in the first place. Furthermore, it is known that a genome-wide gene-environment interaction test has better power than a vqtl prioritization approach when the E is known. (Pare et al, Plos Genet 2010), however, iterating through multiple environments immediately and harshly diminish power due to multiple testing. I would argue only findings that surpass this more stringent threshold be presented, or at the very least, some clearer communication about this rationale.
4. Similar to #3, I find the replication P-value correction insufficient. The authors define significant by the number of SNPs tested in each cohort, but in reality, they are testing 98 total SNPs for replication (no matter which cohort).
5. It is most appropriate to only discuss loci that surpass both a strict study-wide significant threshold and replicate (at least in the same biobank). Therefore, I recommend at least mentioning these limitations (replication and p-value) throughout the manuscript when loci that did not replicate are highlighted. Furthermore, more discussion around the lack of replication is warranted.
6. Unfortunately, the GWAS catalog is not a reliable resource for classifying loci or trait-loci pairs in a GxE study as known versus novel given so few of these findings have made it to that resource. A quick check, and I found at least 7 trait-gene pairs that have already been identified in the same exact dataset, UK Biobank (APOE-lipids, LEPR-CRP, ALPL-ALP, PNPLA3-AST, CRP-CRP, CPS1-Plt, SLC2A9-Urate/Gout; Bernebeu et al Nat Genet 2021, Westerman et al Nat Commun 2022). Many of your signals are driven by Sex and the Bernebeu et al manuscript tested UKB-wide for gene-sex interactions on >500 traits, and similarly, the Westerman et al manuscript tested UKB-wide for GxE with >2000 environmental interactors on blood biomarkers. A more thorough look at what has already been done in UKB or BBJ in the GxE space would give a more accurate answer to the question of novelty. The true novelty of this manuscript is not in the GxE discovery, but rather the deeper investigation at top loci.

7. Provide rationale for dividing the UK Biobank by British European and non-British European participants for discovery and replication. Given the authors are already testing for cross-population replication in two different biobanks/ancestries as their next step, examining consistency and replication within the same population should also be given emphasis. Randomly allocation of participants in UKB into discovery and replication cohorts is more appropriate.
8. Given the possible heterogeneity of the discovery and replication cohorts, it is less surprising that there was lower replication of GxE interactions between UKB1 and UKB2. This should be stated as a limitation in the limitations section.
9. If the cross-population replication was so low, why do the BBJ significant loci consist of mostly shared loci? Does this suggest that there may be more undiscovered loci in the BBJ population? Page 13, Line 248
10. For replication cohorts, authors state that replicated GxE interactions were driven by the same environments as the discovery cohorts (Page 13, Line 240) but do not state this in the cross-population analysis (Lines 253 – 254). Please state whether this is true.
11. Page 14 262 – 263 Authors state that lead variants were not testable in the other population for 19 loci. Why were proxies not used for these loci? If not possible please provide justification.
12. Line 514 – 515, does this imply that PGS should be applied in consistent environmental settings?
13. Could the effect of drinking and the ALDH2 locus on T2D be mediated through hemoglobin levels, and be unrelated to metabolic disease?
14. Do you have interaction effects for natto x PITX2 locus on arrhthmia
15. Line 375, Can you include a significance test demonstrating that the proportion of arrhythmia patients is higher in the homozygous carriers of natto non-consumers? I think figure 2g,h, and i are misleading without this test for comparison. While it does appear that among AA carriers there is a higher rate of individuals that do not eat natto than AC or CC genotypes, there is no difference in the proportion of individuals who do not eat natto in the other plots. The bottom right corner of these figures appears darker simply because that genotype and not eating natto are more common. I am therefore skeptical or unclear about the warfarin findings. Please explain more clearly with statistical tests.
16. Please discuss the possibility of survival bias resulting in age bins that are heterogeneous, therefore explaining some of the age-interaction effects.

Referee #3

(Remarks to the Author)
See comments from other reviewer for this joint review.

Referee #4

(Remarks to the Author)
Understanding how genetic variant effects on health outcomes differ by environmental contexts is an important and timely area of research which can address emerging needs and gaps in the field. An incredibly large number of analyses and enormous scope of work was conducted, using 2 large biobank samples consisting of over 500,000 participants, to generate a resource of GxE interactions for the scientific community and provide exemplar GxE insights into biologic mechanisms for health outcomes.

The authors performed genome-wide gene-environment interaction (GxE) analyses for 47 phenotypes and 12 environments, identifying 92 GxE at 53 loci (several of which have been previously reported) in total. Additional GxE analyses were performed for just under 3,000 protein and 325 metabolite measures. Interestingly, they report molecular changes (metabolome/proteome) related to loci identified in the GxE scan, suggesting biologic pathways involved in gene-environment interplay. Finally, pleiotropic G×E interactions were tested for another 87 diseases across the 2 biobanks, GxE heritability and genetic correlation and differences in polygenic prediction accuracy in different environmental contexts was assessed. The authors report that most of the pleiotropic GxE interactions occurred within a set of related traits and were not commonly observed across phenotypic trait categories. Single cell analysis revealed cell types and biology relevant to the observed differences in genetic effects on pulse pressure by age strata.

Overall, this work is highly significant and I found the methods employed to be rigorous and thorough. Additional strengths of the paper include discovery and replication samples and across 2 populations, “re”discovery of well known GxE and identification of new GxE and considerable secondary analyses to highlight the biologic relevance of these findings. I do have some concerns below that should be addressed.

1) The shaping of the introduction and utility of GxE results for precision medicine purposes is important and makes sense. What I struggle with a bit is that all of the analyses and plots and interpretations we can make are at the population level. There seems to be considerable heterogeneity at the individual participant level – how informative will the results really be for precision medicine at an individual level? Can these results really predict on an individual level?

1) The UKBB is not representative of the population, given it is largely comprised of healthy volunteers, which has been shown to influence GWAS estimates: https://doi.org/10.1038/s41562-023-01579-9. This should be recognized in the discussion including whether the authors think this is likely to influence their results.

2) Clustering dietary survey responses to define exposure groups/clusters. How are those meaningful to interpretation? For example, how can you use the results to provide a recommendation for reducing or increasing consumption of a dietary product in the context of genetic background if you don’t know what amount/level of consumption matters because the measure of exposure you detected was a group based assignment of individuals with similar questionnaire/survey response profiles (principal components) instead of the value of the exposure itself? And could the environment clusters/groups be

biased/clustering by some other factor related to how they answered the questionnaire?

3) The results report some loci show GxE (e.g. ALDH2) associations with multiple environments and argue this provides biologic mechanism insights. I agree biologic insights can be gained but is it also possible some of the same GxE across multiple exposures could be due to exposure correlation?

4) Are these clinically meaningful differences in the effects of E by G? They reach statistical significance but what is the absolute or relative risk difference by G? For example, in Supplementary Figure 4, there seems to be a very small difference in eGFR change by genotype even in the oldest age, around 70 years.

5) The results section on “A disease-induced dietary change caused a GxE interaction” is particularly important, I think, to highlighting the causality and directionality of GxE is not always upstream of the outcome. This may be something that the authors want to state in the discussion/highlight as an important consideration with GxE result interpretation, generally. I think many in the field will assume it is the gene-environment interaction causing the outcome and don’t always consider that gene-outcome/treatment can cause people to change their “environment” and lead to a “GxE”.

6) In the introduction section, this sentence isn’t clear to me: “We further projected environment-stratified genetic effects to scRNA-seq data to uncover the dynamic shift of responsible cell types at single-cell resolution”. I don’t really understand what question you sought to ask or how you answered it methodologically from this sentence.

7) In Supplementary Figure 3 – what is the reference population used for the r2 correlation values shown? European? What does the correlation structure look like in other global populations?

8) Figure 3a-b reports GxE heritability metrics. It would be helpful to also see (for reference) what the heritability estimate for G and E alone would look like in comparison to these GxE estimates. I struggled a bit to understand the point here and how it would relate to my expectations of more “traditional” heritability estimates for outcomes assessing marginal G or E effects.

9) The PGS and cross-population portability fundamentally seems like a different question and scope than the rest of the paper detecting locus-specific GxE and biologic mechanisms. Could be a separate and equally impactful brief report. As is, it gets lost in the length of the results and is important to the field!

10)The single cell PP follow up result also seems specific to a condition and its own important discovery but if length not an issue okay to keep in the paper.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
My previous concerns have been fully addressed by the authors. I have only one minor comment remaining:

Lines 485–498 – Reference for G×E-PGS:
In the section beginning “Consistent with these findings, PGS constructed from G×E interactions (G×E-PGS) significantly...”, the authors note that elastic net was used for the G×E-PGS analyses. However, the model itself—comprising PGS_additive, PGS_G×E, and the interaction term PGS_G×E × E—is conceptually based on the G×Eprs framework. As such, the G×Eprs model should be properly referenced to acknowledge that this modeling approach is pre-existing and has already been validated. While the supplementary methods provide some explanation, the manuscript would benefit from a citation to an established reference, if available.

If, on the other hand, G×E-PGS is intended to represent a novel methodological contribution, this should be clearly stated in the manuscript. In that case, it would also be important to support the approach with simulations or additional analyses to demonstrate its validity and robustness.

Referee #2

(Remarks to the Author)
I greatly appreciate the author’s thorough response to reviewer comments. In nearly all areas, I am satisfied with their responses. I particularly appreciate removing some over-stating claims and incorporating additional cohorts and longitudinal data. I still have 1 major and 3 minor comments
Major comment
1. Mendelian randomization (MR) - While MR can be a powerful approach for assessing causal relationships, its assumptions are frequently violated, especially in studies of environmental exposures. In this context, the validity of the genetic instruments is particularly questionable due to the high likelihood of horizontal pleiotropy.
The authors rely solely on the inverse-variance weighted (IVW) method, which assumes that all genetic instruments are valid and that there is no directional pleiotropy. However, even a single pleiotropic variant can bias the results. No sensitivity

analyses (e.g., MR-Egger, weighted median, or heterogeneity tests) were performed to assess pleiotropy or potential violations of MR assumptions.
Regarding the Natto/warfarin example, I believe temporal ordering alone provides sufficient evidence for the reverse causal relationship, and the MR analysis adds little value—particularly without robust validation of the instruments. Unless the authors can convincingly justify that the genetic instrument for Natto consumption is valid and not affected by horizontal pleiotropy, I recommend removing this MR analysis.
I also encourage the authors to consult the MR-STROBE guidelines and include a checklist if MR is retained in any part of the manuscript. The recent preprint by Sutton et al. (https://doi.org/10.1101/2025.06.26.25330002) provides an overview of the concerns in conducting MR with dietary exposures and reinforces the need for rigorous instrument validation.
Overall, I recommend removing the MR component from the manuscript unless a more in-depth analysis is performed, with appropriate sensitivity checks, justification for instrument validity, and MR STROBE checklist (https://www.strobe-mr.org/).
An additional concern that I have with retaining it, is that this publication is already dense, and overly-detailed with a battery of tests. Any more detail in MR would further decrease the readability of the manuscript
Minor comments
1. In Line 462 you discuss the lack of correlation between heritability of vQTLs and GxE interactions is not necessarily surprising, given your GxE interaction analyses of only a small subset of exposures is expected to capture all the underlying variance, which could be due to many other GxE and GxG. I think your interpretation of this result is not accurate - that vQTL failed to fully capture GxE interactions. Please modify this.
2. Although the results catalog is more comprehensive than prior studies, the manuscript still overstates its scope by referring to the analysis as "phenome-wide" or "exposome-wide" when in fact only a limited number of phenotypes and exposures were tested. I recommend softening these claims to better reflect the study's actual breadth.
3. The manuscript remains highly information-dense, to the point of being difficult to read. I encourage the authors to reduce the number of details in both the text and figures wherever possible. Streamlining the presentation would improve readability.

Referee #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks to the Author)
Thank you for careful consideration of the initial comments provided. The revisions strengthened aspects of the prior version that were concerns for the reviewers and I have no further concerns at this time.

Response to the reviewer's Comments:

Reviewer #1:

A Cross-population Compendium of Gene–Environment Interactions
Shinichi Namba et al.

This study explores gene–environment (G×E) interactions using large biobank data from 543,732 European and Japanese individuals, covering 47 phenotypes and 12 environmental factors. It provides a global overview of G×E interactions, focusing on replication, cross-population consistency, and the pleiotropic effects of genes on disease risks. The study highlights the role of G×E interactions in understanding phenotypic plasticity, identifying how age, sex, and lifestyle factors contribute to specific loci, including a notable example of reverse-causal G×E interaction related to disease-driven dietary changes. Through genome-wide interaction studies, the authors address the challenge of missing heritability within the G×E framework. They also illustrate the influence of G×E interactions on polygenic prediction accuracy and cross-population applicability. The authors contend that this study serves as a vital resource for understanding the dynamics of genetic associations, with significant implications for complex trait biology, personalized medicine, and drug development.

This is an important and well-written study that provides a comprehensive analysis. While the manuscript is strong overall, I have a few comments that may be critical to address. Incorporating these suggestions could significantly enhance the clarity and impact of the current version.

Response:

We sincerely thank you for taking your time to consider our manuscript and for kindly acknowledging the value of our manuscript. We appreciate your insightful comments to improve the manuscript further. Your suggestions to evaluate the causal directions underlying the G×E interactions were particularly helpful. In the point-by-point response below, we provide additional analyses, including Mendelian randomization and G×E tests conducted using data arranged in a temporal order from the survey of environmental factors to the assessment of clinical phenotypes. To improve readability, the changes made to the figures are highlighted with red boxes.

1. Causal Directions: It remains unclear how the authors establish the causal direction in their analysis. For instance, could the genetic effects on alcohol consumption (such as rs672's impact on drinking behavior) be influenced by factors like AST or Hb levels? The fact that drinking behavior was not measured prior to assessing AST or Hb levels presents a potential limitation in the study design. Or, please clarify if this is not the case. This limitation arises from the inability to establish clear temporal causality—specifically, whether drinking behavior influenced AST and Hb levels or whether pre-existing conditions or genetic predispositions (such as rs671) may have influenced both alcohol consumption and the biomarkers. To

resolve this, the authors might consider conducting a Mendelian Randomization (MR) study to establish a causal relationship and avoid ambiguity around the directionality of effects.

Response:

Thank you for your thoughtful suggestion. We agree with you that the temporal order between environments and phenotypes is helpful to determine the causal directions. To systematically test the causal directions of G×E interactions, **we newly leveraged repeat assessment data for biomarkers to ensure that the environmental values were measured prior to assessing phenotypes. For disease outcomes, we newly performed Cox regression analyses focusing on incidental cases with disease onsets at least six months after baseline, thereby confirming that the environmental exposures preceded disease development.**

First, for the biomarkers, we confirmed significant G×E interactions for 15 / 31 of biomarker-locus pairs in BBJ1 (N_{max} = 105,339). In an independent Japanese cohort, J-MICC/HERPACC, the second measurement data were available (N_{max} = 37,398), which were assessed five years after the baseline assessment. In this cohort, the second measurement data were available for 27 G×E interactions, and 11 were significantly verified. Notably, the G×E interactions at the *ALDH2* locus, which exhibited high pleiotropy, were consistently verified in both BBJ1 and J-MICC/HERPACC. Particularly, the G×E interactions for AST and Hb were significantly verified in both BBJ1 ($P_{G\times E}$ = 1.4×10^{-11} and 1.1×10^{-17}) and J-MICC/HERPACC ($P_{G\times E}$ = 1.6×10^{-11} and 1.1×10^{-12}).

For the European population, we primarily analyzed All of Us rather than UKB1, as described below. In UKB1, despite the availability of second to fourth follow-up visits (N_{max} = 50,958), the G×E interactions were not significantly verified. This result is likely attributable to the biased characteristics reported for the participants of the repeat visits in UKB, who tended to be older, less socio-economically depleted, and never smokers, with higher education and excellent overall health rating (https://biobank.ndph.ox.ac.uk/~bbdatan/repeat_assessment_characteristics.pdf). We repeated the analysis using baseline measurements from the same subset of participants, and again, the G×E interactions were not significantly verified, suggesting that the selection bias, rather than measurement timing, was a key limiting factor of validation in UKB1. Therefore, we performed the assessment using the European population in All of Us (N_{max} = 47,292), and 4 out of 60 testable G×E interactions were significantly verified. These results support the causal direction from environments to clinical traits for the verified G×E interactions.

Next, for the disease traits, the G×E interaction for T2D at the *ALDH2* locus, primarily driven by G×Ever-drinking, was significantly verified in BBJ1 ($P_{G\times E}$ = 5.3×10^{-5}). In UKB1, no G×E interaction was significantly verified, while the G×Ever-drinking interaction for arrhythmia at the *STARD3NL* locus was nominally verified ($P_{G\times E}$ = 0.026).

We further assessed the temporal directions by conducting Cox regression analyses for overall survival, where the temporal direction from environment to death is unambiguous. We identified a G×E interaction at the *ALDH2* locus in BBJ1 driven by interactions with sex, ever-drinking, and age ($P_{G\times E}$ = 1.7×10^{-11}), indicating that the G×E interaction may affect human lifespan.

We believe that these additional analyses have clarified the temporal order between environmental exposures and outcomes, thereby improving the interpretability of our findings. We sincerely appreciate your suggestion. However, we acknowledge a limitation in using

temporal order to assess the causal direction of G×E interactions: both measurement values and environmental exposures can be temporally stable and may change only gradually over time, making it difficult to strictly infer causal direction based solely on the timing of measurement. Moreover, sex and age are inherently unaffected by clinical traits. In some cases, clinical traits and environmental exposures may arise simultaneously, or G×E signals may appear statistically without a direct biological mechanism (Westerman KE and Sofer T, ***AJHG*** (2024)). Therefore, while we agree with you that clarifying the causal direction can help yield insights into the underlying mechanisms of some G×E interactions, this may not be the case for all such interactions.

We further performed bidirectional Mendelian randomization (MR) analyses using the inverse-variance weighted method, which revealed causal associations from BMI to ever-smoking and current-smoking in UKB (P = 1.7×10^{-4} and 2.7×10^{-5}, respectively). These associations might underlie the G×Ever-smoking and G×Current-smoking interactions observed at the *FTO* and *HYKK* loci, respectively. We also evaluated the temporal order of the G×Natto interaction at the *PITX2* locus and conducted MR analyses, as detailed in our response to your comment #2.

Modification:

We added the results of temporal order analyses and MR in **Results** and **Supplementary Note 8** as follows.

"To help distinguish causal directions between environments and clinical traits for other loci, we leveraged the repeat measurement data for the biomarkers to sort out the temporal order from environments to phenotype measurements, performed time-to-event Cox analyses for the disease statuses, and conducted bidirectional Mendelian randomization (**Supplementary Note 8; Supplementary Tables 17 and 18**). We also performed Cox analyses for overall survival and identified a significant G×E interaction at the *ALDH2* locus driven by interactions with sex, ever-drinking, and age (P = 1.7×10^{-11}), suggesting that the G×E interaction may affect human lifespan." (**Results**, P16–17 line 383–390).

"**Supplementary Note 8. Potential causal directions between environments and clinical traits**

The discovery of the reverse causal relationship of the G×Natto interaction at the *PITX2* locus prompted us to a broader investigation into the causal directions of other G×E interactions identified in this manuscript. Specifically, we examined whether environmental factors influenced clinical traits by leveraging longitudinal data from multiple cohorts.

For the biomarkers, we leveraged the most recent repeat measurement data and restricted the data to those measured at least six months after the baseline survey (**Supplementary Table 17**). In BBJ1, the median time interval from the baseline to the phenotype assessment was 4.0 years (range, 0.5–9.5; N_{max} = 105,339). Among 31 trait–locus pairs, 15 were significantly verified using the repeated measurement data ($P_{G\times E} < 0.05/98 = 5.1\times10^{-4}$; **Supplementary Table 18**). In J-MICC/HERPACC, the second measurement data were available (N_{max} = 37,398), which were assessed five years after the baseline assessment. In this cohort, 11 out of 27 testable G×E interactions were significantly verified. Notably, the highly pleiotropic G×E interactions at the *ALDH2* locus were consistently verified in both BBJ1 and J-MICC/HERPACC.

For the European population, we primarily analyzed All of Us rather than UKB1, as described below. In UKB1, the second to fourth follow-up visits were available (median time interval = 5.0 years [range, 2.0–17.0]; N_{max} = 50,958). Nevertheless, the G×E interactions were not significantly verified. This result is likely attributable to the biased characteristics reported for the participants of the repeat visits in UKB, who tended to be older, less socio-economically depleted, and never smokers, with higher education and excellent overall health rating (https://biobank.ndph.ox.ac.uk/~bbdatan/repeat_assessment_characteristics.pdf). We repeated the analysis using baseline measurements data from the same subset of participants. The G×E interactions were again not significantly verified, suggesting that the selection bias, rather than measurement timing, was a key limiting factor of longitudinal validation in UKB1. Therefore, we performed the assessment using the European population in All of Us (median time interval = 2.2 years [range, 0.5–6.3]; N_{max} = 47,292). In this cohort, 4 out of 60 testable G×E interactions were significantly verified (**Supplementary Table 18**). These results support the causal direction from environments to clinical traits for the verified G×E interactions.

As a limitation of this approach, we note that the measurement values and environments can be temporally stable and may change only gradually over time, which complicates the strict inference of causal direction based solely on measurement timing. Moreover, sex and age were inherently unaffected by clinical traits, and therefore, the repeat assessment analysis may not serve as additional evidence for the causal direction of these factors. In some cases, clinical traits and environmental exposures may arise simultaneously, or G×E signals may emerge statistically without a direct biological mechanism[6].

For the disease traits, we focused on incidental cases, whose disease onsets were at least six months after the baseline assessment, and performed Cox regression with G×E interaction terms. In BBJ1, the G×E interaction for T2D at the *ALDH2* locus, primarily driven by G×Ever-drinking, was significantly verified ($P_{G\times E} = 5.3\times10^{-5}$). On the other hand, the G×E interaction for arrhythmia at the *PITX2* locus was not significant ($P_{G\times E} = 0.077$). In the section “A disease-induced dietary change caused a G×E interaction,” we discussed that this locus was driven by a reverse causal relationship from a disease to an environment (specifically, from warfarin medication for atrial fibrillation/flutter to the avoidance of natto consumption). Considering that the Cox regression model tested the temporal direction from environments to diseases, the failure of its verification in the Cox regression was expected. In UKB1, no G×E interaction was significantly verified, while the G×Ever-drinking interaction for arrhythmia at the *STARD3NL* locus was nominally verified ($P_{G\times E} = 0.026$).

To assess the causal direction more confidently, we performed Cox regression for overall survival, where the temporal order from environmental factors to the endpoint of death was apparent. Testing 53 loci in their discovered cohorts, we observed a significant G×E interaction for the *ALDH2* locus in BBJ1 ($P_{G\times E} = 1.7\times10^{-11}$; driven by G×Sex, G×Ever-drinking, and G×Age), indicating that the G×E interaction may affect human lifespan. In UKB1, we observed a significant G×Age interaction at the *APOE* locus ($P_{G\times E} = 5.5\times10^{-8}$), consistent with its G×Age interactions for total cholesterol (TC) and triglycerides (TG). As its effect size was in the same direction as the additive genetic effect (0.098 [95% confidence interval, 0.065–0.130] and 0.054 [0.020–0.087], respectively), this interaction suggested a larger mortality risk for carriers of the risk variant than non-carriers at older ages.

We also performed bidirectional Mendelian randomization analyses using the inverse-variance weighted method to complement the analyses using longitudinal data (**Supplementary Methods**). We observed significant causal associations from BMI to ever-smoking and current-smoking in UKB (P = 1.7×10^{-4} and 2.7×10^{-5}, respectively). These associations might underlie the G×Ever-smoking and G×Current-smoking interactions observed at the *FTO* and *HYKK* loci, respectively.

Collectively, we systematically investigated the causal directions between environments and phenotypes, which may help future studies to elucidate the mechanisms underlying these G×E interactions." (**Supplementary Note 8**).

"For BBJ, we used the biomarker phenotypes measured at the nearest dates to the baseline assessment for the main analyses, while we used those measured most recently for the temporal order analyses. For UKB, we used the baseline assessment data for the main analyses, while we used the most recent data from the second to fourth revisit assessments for the temporal order analyses, namely, the first repeat assessment visit, the imaging visit, and the first repeat imaging visit. We applied the same quality controls for both biobanks, including (i) excluding participants with age <18 or age >85, (ii) excluding participants with particular disease status that might have affected the phenotype values, (iii) correction of phenotype values for participants taking anti-hypertensive medications or statins, (iv) excluding outliers whose measured values were outside three times of interquartile range of upper/lower quartile or outside three times of standard deviation from mean, and (v) applying natural log transformation for the phenotypes with right-skewed distributions. For the temporal order analyses, we restricted the data to those measured at least half a year after the baseline survey." (**Methods**, P39–40 line 988–1001).

"All statistical tests were two-sided unless otherwise noted. As phenotypes, we targeted quantitative traits (clinical biomarkers, metabolites, and protein expression measurements), dichotomous traits (disease statuses), and time-to-event traits (disease onset for incidental cases and overall survival). For quantitative and dichotomous traits, we tested G×E interactions using GEM, a fast and scalable implementation of linear and logistic regressions with G×E interaction terms[23] (**Ext. Data Fig. 1**). We estimated the model-robust "sandwich" standard errors to suppress the inflation of statistics[23]. As the case–control imbalance can cause the inflation of test statistics in the logistic regression, we re-estimated the effect sizes and P-values using Firth logistic regression for the variants with P-values less than 5.0×10^{-8} for dichotomous traits. For the time-to-event traits, we employed the Cox proportional hazard model implemented in the R package "survival". Individuals who had already developed the target disease at baseline or developed it within six months from baseline were excluded from the incidental case analyses." (**Methods**, P43 line 1077–1089).

"**Mendelian randomization**

To assess the causal direction between environmental factors and clinical traits, we performed bidirectional Mendelian randomization, where we performed two Mendelian randomization analyses using (i) environmental factors as exposure and clinical traits as outcome, and (ii) environmental factors as outcome and clinical traits as exposure. We tested

all pairs of clinical traits and the environments contributing to the genome-wide significant G×E interactions.

To obtain GWAS summary statistics used for Mendelian randomization, we first performed GWAS in BBJ1, BBJ2, UKB1, and UKB2 using PLINK2. We applied linear regression models for continuous traits and Firth hybrid logistic regression models for dichotomous traits. We used the lead variants of exposure traits as instrumental variables except for those in the *ALDH2* locus, considering that the highly pleiotropic nature of this locus might violate the assumption of Mendelian randomization that there was no horizontal pleiotropy. We used GWAS summary statistics from BBJ1 and UKB1 for exposure and those from BBJ2 and UKB2 for outcome, respectively, to ensure that there was no sample overlap between the two GWAS summary statistics used for Mendelian randomization.

We employed the inverse-variance weighted approach for Mendelian randomization. As this approach require multiple instrumental variables, we did not perform Mendelian randomization when at most one locus was detected in GWAS. To meet this requirement, we additionally used publicly available GWAS summary statistics for drinks per week (DPW) and cigarettes per day (CPD) from our previous studies. As these GWAS studies included BBJ1 in the discovery cohorts and did not include BBJ2, we used them for exposures and the summary statistics from BBJ2 for outcomes. We conducted all Mendelian randomization analyses using the R package TwoSampleMR." (**Supplementary Methods**).

2. Reverse Causality: The authors mention, "In summary, we identified the G×E interaction caused by a reverse causal relationship from the disease to the environment" (line 393). However, while the study suggests reverse causality—such as atrial fibrillation (AF) leading to changes in natto consumption due to warfarin treatment—the evidence is suggestive rather than conclusive. Several alternative mechanisms, such as confounding factors, pleiotropy, incomplete data on treatment timing, or non-dietary factors, could also explain the observed G×E interaction. To strengthen their findings and rule out alternative explanations, the authors could consider applying a causal inference analysis, such as MR or other techniques, to clarify the causal pathways and ensure robust conclusions.

Response:

Thank you for your thoughtful comment. We agree with you that there were additional contributors at the *PITX2* locus, as we observed that G×Sex, G×Ever-drinking, and G×Ever-smoking also contributed to this locus. Nevertheless, **testing the G×Natto interaction alone surpassed the genome-wide significant threshold ($P_{G\times E}$ = 2.1×10^{-10})**, supporting the primary contribution of G×Natto interaction at this locus. To directly observe the causal effect between warfarin medication and natto consumption on treatment timing, **we newly compared the frequency of natto consumption in the same individuals before and after warfarin initiation. We observed a marked decrease in natto consumption after warfarin initiation** (**Fig. 2i**), providing temporal evidence consistent with a reverse causal direction from disease (atrial fibrillation) to environment (natto avoidance due to anticoagulant treatment). We also conducted an MR analysis and observed a significant causal association from warfarin medication to decreased natto consumption (P = 6.6×10^{-3}). A Cox regression analysis using incidental cases in BBJ1 was not significant at this locus ($P_{G\times E}$ = 0.077), which was expected given that the Cox regression model tested the temporal direction from environments to diseases, opposite to the temporal order observed above.

Furthermore, **we found that the marginal effect size of the lead variant was heterogeneous between the subgroups of arrhythmia**, making the overall effect size for arrhythmia dependent on the rate of atrial fibrillation/flutter patients among the arrhythmia patients. This led to a difference in the effect size for arrhythmia across strata of natto consumption frequency, which explained the link between the reverse causality and the G×Natto interaction. In line with this notion, the G×Natto interaction was not significant when evaluated separately within the atrial fibrillation/flutter and the other arrhythmia subgroups ($P_{G\times E}$ = 0.84 and 2.6×10^{-4}, respectively).

Together, these results support the reverse causal relationship from the disease to the environment, consistent with clinical practice in Japan, where physicians commonly advise warfarin-taking patients to avoid eating natto.

Modification:

We added the additional evidence to support the reverse causal direction in **Results**.

"Testing G×Natto alone surpassed genome-wide significance ($P_{G\times E} = 2.1\times10^{-10}$), supporting its primary role at this locus." (**Results**, P15 line 350–351).

"Indeed, natto intake declined markedly after warfarin initiation in the same individuals (**Figure 2i**). Mendelian randomization supported the negative causal effect of warfarin on natto consumption ($P = 6.6\times10^{-3}$, effect size = -0.029 [95% confidence interval (CI), -0.051 to -8.2×10^{-3}]).

When stratifying arrhythmia patients by clinical subgroups, the marginal effect size of rs72900155 was substantially larger for atrial fibrillation/flutter than for other subgroups (0.49 [95% CI, 0.46–0.52] versus 0.13 [0.10–0.15]). Due to this effect size heterogeneity, the overall effect size for arrhythmia would depend on the proportion of atrial fibrillation/flutter patients and therefore differ across natto intake strata, explaining the link between the reverse causality and the G×Natto interaction. Consistently, the G×Natto interaction was not significant within either subgroup when evaluated separately ($P_{G\times E}$ = 0.84 for atrial fibrillation/flutter; 2.6×10^{-4} for other subgroups)." (**Results**, P16 line 360–371).

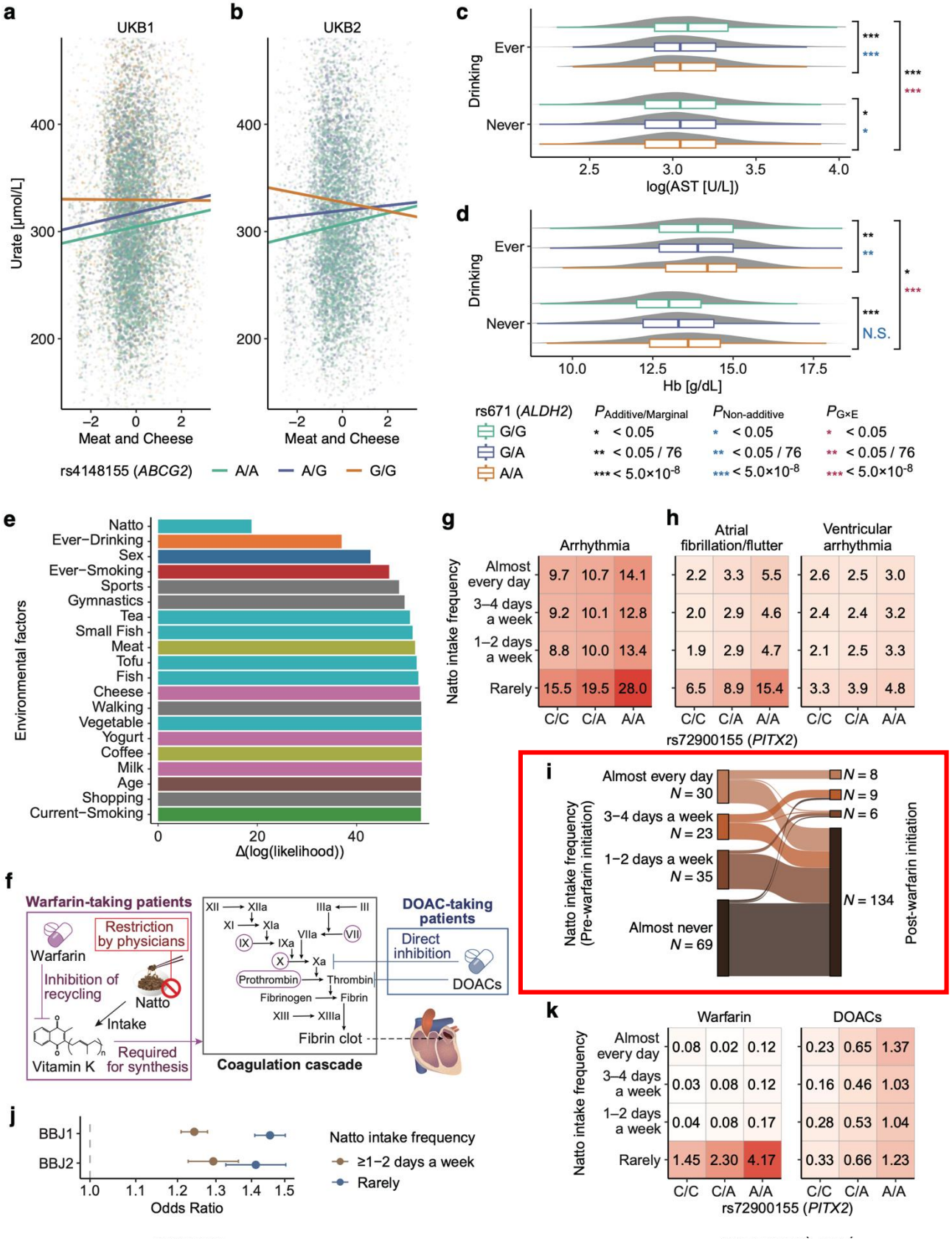


Figure 2, Representative loci with G×E interactions.

a and **b**, Urate levels across rs4148155 genotypes (the *ABCG2* locus) and "meat and cheese" consumption in UKB1 (**a**) and UKB2 (**b**). Dots show 5,000 randomly sampled individual measurements per genotype; lines represent the linear regression lines using all participants. **c** and **d**, AST (**c**) and hemoglobin (Hb, **d**) levels by rs671 genotypes (the *ALDH2* locus) and drinking status in BBJ1. Boxplots show median, quartiles, and 1.5× interquartile range whiskers. **e**, Log-likelihood improvement for G×E models of rs72900155 (the *PITX2* locus) for arrhythmia. Environmental factors were added stepwise to the null model (**Methods**). **f**, schematic of the mechanisms of action for warfarin, natto (fermented soybean), and direct oral anticoagulants (DOAC). Heart icon from TogoTV (© 2016 DBCLS TogoTV, CC-BY-4.0). **g**, heatmap of the arrhythmia prevalence in BBJ1. **h**, same as **g**, stratified by arrhythmia subgroups. **i**, natto intake before and after warfarin initiation in BBJ atrial fibrillation/flutter patients with ≥1 PT-INR record per year, considering that regular monitoring of PT-INR is required during warfarin therapy. **j**, odds ratios of rs72900155 (the *PITX2* locus) for arrhythmia stratified the natto intake frequency. Error bars represent 95% confidence intervals (CI; Firth logistic regression). **k**, heatmap of the atrial fibrillation/flutter prevalence in BBJ2, stratified by warfarin and DOAC medication.

We also described the Cox regression analysis using incidental cases in **Supplementary Note 8**.

"On the other hand, the G×E interaction for arrhythmia at the *PITX2* locus was not significant ($P_{G\times E}$ = 0.077). In the section "A disease-induced dietary change caused a G×E interaction," we discussed that this locus was driven by a reverse causal relationship from a disease to an environment (specifically, from warfarin medication for atrial fibrillation/flutter to the avoidance of natto consumption). Considering that the Cox regression model tested the temporal direction from environments to diseases, the failure of its verification in the Cox regression was expected." (**Supplementary Note 8**).

3. There is a risk of generating spurious signals in G×E studies when adjusting for covariates, particularly due to collider bias. I notice that multiple covariates were used in the analysis. How do the authors ensure that the covariates selected do not introduce spurious signals? A detailed explanation of the covariate selection process, particularly how potential colliders were handled, would be beneficial.

Response:

Thank you for your important comment. We used the following covariates: genetic principal components (PCs), age × sex, age^2, and age^2 × sex. We also adjusted for genotyping array in UKB. We also adjusted for the disease statuses used as inclusion criteria at recruitment for quantitative traits and overall survival in BBJ, following previous large-scale GWAS using BBJ (Kanai M *et al.* ***Nat Genet*** (2018), Sakaue S *et al.* ***Nat Genet*** (2021), Wei Z *et al.* ***Cell Genom*** (2023)). For the metabolome, we also adjusted for statin usage, considering its strong effect on metabolite levels. **We note that these covariates were available for almost all participants.** In addition to these covariates, the environmental factors were also used as covariates in principle in G×E interaction tests. To minimize potential collider bias due to non-random missingness of environmental factors, we selected questionnaires on dietary and physical activity lifestyles, considering their availability in both biobanks and their low rates of

missing data. Indeed, the missing rates of the environmental factors for lifestyle were limited to 0.01–0.27% in UKB1 and 1.4–13.3% in BBJ1.

As you thoughtfully suggested in #4, replication would support the reliability of the detected G×E interactions. The replication rates of G×E interactions were 11.3% (7/62) in UKB2 and 52.8% (19/36) in BBJ2, at the study-wide Bonferroni-corrected significance threshold. The rate was higher in BBJ2, possibly due to its larger number of participants. To empirically evaluate these rates in comparison to those of GWAS, we performed GWAS for the traits with genome-wide significant G×E interactions in UKB1 and BBJ1. We used the same covariates for GWAS as for G×E interaction tests. As environmental factors were not included in the GWAS model, their potential collider bias did not affect the GWAS signals in principle. When using all GWAS loci that surpassed the Bonferroni-corrected threshold, the replication rates were 10.0% (522/5,238) in UKB2 and 34.0% (613/1,802) in BBJ2, which were lower than those of G×E interactions. In this analysis, the Bonferroni-corrected thresholds were different between G×E and GWAS. Therefore, we next filtered the GWAS trait–locus pairs to match the Bonferroni-corrected thresholds. Specifically, for each G×E locus, we selected the marginal effect locus for the same trait whose *P*-value was closest to that of the corresponding G×E interaction. Using this matched set, the replication rates of GWAS were only slightly higher than those of G×E (14.5% (9 / 62) in UKB2 and 66.7% (24 / 36) in BBJ2). In summary, these results demonstrate that **the replication rates of G×E interactions reported in our manuscript were largely comparable to those of GWAS**. Considering that differences in environmental distributions across cohorts can reduce the replicability of G×E interactions, and that previous G×E studies have been criticized by their low replication rates, these findings suggest the high reliability of the G×E interactions identified in this manuscript.

We then tested for replication in the newly added four independent cohorts with diverse populations, following your thoughtful suggestion in #4 (All of Us, J-MICC/HERPACC, JPHC, and HPP; comprising European, East Asian, African, American, and Israeli populations; *N* = 436,272). We note that although BBJ2 belonged to the same biobank as BBJ1, it was nearly independent of BBJ1, as most participants (83%) were recruited separately from BBJ1. The completely independent cohorts included the European population in All of Us (N_{max} = 208,700; in the US) and the East Asian population in J-MICC/HERPACC (N_{max} = 70,909; in Japan) for replicating the UKB1 and BBJ1 results, respectively. In these cohorts, 27% (17/62) and 56% (20/36) trait–locus pairs were significantly replicated ($P_{G\times E} < 0.05/(62+36) = 5.1\times10^{-4}$), respectively. Notably, the pleiotropic G×E interactions at the *ALDH2* locus showed a particularly high replication rate in J-MICC/HERPACC (81% = 17/21). We also tested replication in another Japanese cohort, JPHC (N_{max} = 10,904), where 6 G×E interactions at the *ALDH2* locus were again significantly replicated.

Furthermore, we used the African and American populations in All of Us (N_{max} = 70,558 and 66,556, respectively) and the Israeli population in HPP (N_{max} = 8,645; largely composed of Ashkenazi Jewish people) to evaluate cross-population replication in populations other than the European and East Asian populations. In the African population, three G×E interactions were significantly replicated, and two of them were included in the shared G×E interactions between UKB1 and BBJ1. Two G×E interactions for pulse pressure were significantly replicated in the American population, both of which were originally detected in UKB1 and primarily driven by G×Age in both UKB1 and All of Us. G×E interactions were not significantly replicated in the Israeli population, possibly due to its relatively small sample size. Nevertheless, we observed the top signal for one of the shared G×E interactions between UKB1 and BBJ1 (*UGT1A* gene family for total bilirubin, commonly driven by G×Current-

smoking and G×Age), suggesting its consistency across diverse populations. We summarized the replication statuses, including nominal replication with $P_{G\times E}$ less than 0.05, in **Ext. Data Fig. 5a and b**. We note that we stringently verified that all replicated G×E interactions were driven by the same environments as the discovery cohorts, and the effect sizes were in the same direction. We regarded any G×E interactions that did not meet these criteria as non-replicated, regardless of their *P*-values.

Among the independent cohorts, All of Us is a nationwide cohort, and J-MICC/HERPACC and JPHC are population-based cohorts; the participants of HPP were recruited from volunteers in Israel. **As these cohorts represent diverse populations with different recruitment strategies, questionnaires, and healthcare systems, replication in them supports the generalizability of our findings across heterogeneous settings**.

Modification:

We revised the **Methods** to provide a clearer description of the covariates used in our analyses.

"**G×E interaction testing**

All statistical tests were two-sided unless otherwise noted. As phenotypes, we targeted quantitative traits (clinical biomarkers, metabolites, and protein expression measurements), dichotomous traits (disease statuses), and time-to-event traits (disease onset for incidental cases and overall survival). For quantitative and dichotomous traits, we tested G×E interactions using GEM, a fast and scalable implementation of linear and logistic regressions with G×E interaction terms[23] (**Ext. Data Fig. 1**). We estimated the model-robust "sandwich" standard errors to suppress the inflation of statistics[23]. As the case–control imbalance can cause the inflation of test statistics in the logistic regression, we re-estimated the effect sizes and *P*-values using Firth logistic regression for the variants with *P*-values less than 5.0×10^{-8} for dichotomous traits. For the time-to-event traits, we employed the Cox proportional hazard model implemented in the R package "survival". Individuals who had already developed the target disease at baseline or developed it within six months from baseline were excluded from the incidental case analyses.

We tested G×E interactions for different sets of environmental factors, as shown in **Ext. Data Fig. 1**, and then aggregated *P*-values across environmental sets on the Cauchy distribution[31] to obtain per-variant *P*-values. When *P*-values were 0, which meant *P*-values were less than 1.0×10^{-300}, we estimated the precise *P*-values using the multiple-precision floating-point arithmetic with the maximal precision of 10,000 bits, implemented in the Rmpfr R package. We defined a G×E interaction as significant at the variant level if the aggregated *P*-value was below the genome-wide threshold of 5.0×10^{-8}. We also reported the study-wide Bonferroni threshold for full transparency. We did not require that any individual environment's interaction term be significant on its own, considering that G×E interactions can be driven by multiple environments. Nevertheless, we note that at least one raw *P*-value was smaller than the Cauchy-combined *P*-value in principle, as the Cauchy combination method is not a meta-analysis but rather returns a *P*-value within the range of raw *P*-values.

After evaluating the significance of G×E interactions at the variant level, we determined the order of importance of G×E interaction terms by backward elimination from the regression model with all G×E interaction terms. Specifically, we repeatedly removed the G×E interaction term with the least improvement in the likelihood of the linear regression model or the largest

Wald *P*-value of the Firth logistic regression model. After removing all G×E interaction terms, we brought back the G×E interaction terms one by one in the order of their importance if the model fit was improved with a *P*-value less than 0.05 (likelihood ratio test). We considered that the environmental factors contributed to G×E interactions if the corresponding interaction terms were included in the final model. In principle, this approach selects more parsimonious sets of environments than the Akaike information criterion and can determine the top environment contributing to individual G×E interactions.

We used the first 20 genetic principal components (PC), genotype array, age^2, age × sex, and age^2 × sex as covariates for UKB. For the other cohorts, we used the first 10 PCs, age^2, age × sex, and age^2 × sex as covariates. Age and sex, as well as other environmental factors, were included as the interaction terms with genotypes, and the variables used for interaction terms were automatically included as covariates. We also included the status of the target diseases in BBJ as covariates for quantitative traits and overall survival. We included the status of statin medication as an additional covariate for metabolome measurements. For sex-specific diseases in PheWAS, we excluded sex from the environmental factors and age × sex and age^2 × sex from covariates.

Following that previous cross-population GWAS used the distance-based locus definition[22], we defined that the genome-wide significant variants were in the same locus if their distance was less than 500 kbp. For the genome-wide G×E interaction tests for metabolome, we targeted (i) genotyped variants in each cohort, (ii) HapMap3 variants, and (iii) the variants targeted by the meta-analysis across BBJ and UKB, to reduce the computational cost. After defining the metabolome G×E loci based on their distance as above, we merged the metabolome G×E loci if they overlapped with the same locus defined for the clinical traits." (**Methods**, P43–45 line 1076–1129).

"The questionnaires for dietary consumption and physical activity were available for both UKB and BBJ, with low missing rates (0.01–0.27% in UKB1 and 1.4–13.3% in BBJ1), and were used in this manuscript." (**Methods**, P40 line 1018–1020).

We added the comparison of replication rates between G×E interactions and marginal (GWAS) effects in **Results and Supplementary Note 4**.

"The replication rates were comparable to those in GWAS for the same traits, supporting the robustness of our findings (**Supplementary Note 4**)." (**Results**, P11 line 243–245).

"**Supplementary Note 4. Comparable replication rates of G×E interactions compared to those for GWAS**

To empirically compare the replication rates of G×E interactions to those of GWAS, we performed GWAS for the traits with genome-wide significant G×E interactions in UKB1 and BBJ1. We detected 5,238 and 1,802 trait–locus pairs with genome-wide significant marginal effects in UKB1 and BBJ1, respectively (7,040 in total). In UKB2, 10.0% (522 / 5,238) were significantly replicated at the Bonferroni-corrected threshold of 7.1×10^{-6} (= 0.05 / 7,040), and 34.0% (613 / 1,802) were significantly replicated in BBJ2, reflecting its larger sample size than UKB2. The replication rates of G×E interactions were higher than these rates: 11.3% (7

/ 62) in UKB2 and 52.8% (19 / 36) in BBJ2 at the Bonferroni-corrected threshold of 5.1×10^{-4} (= 0.05 / 98).

In the above analysis, the Bonferroni-corrected thresholds were different between G×E and GWAS. Therefore, we next filtered the GWAS trait–locus pairs to match the Bonferroni-corrected thresholds. Specifically, for each G×E locus, we selected the marginal effect locus for the same trait whose *P*-value was closest to that of the corresponding G×E interaction. Using this matched set, the replication rates of GWAS were 14.5% (9 / 62) in UKB2 and 66.7% (24 / 36) in BBJ2, only slightly higher than those of G×E interactions.

In summary, we observed that the replication rates of G×E interactions were largely comparable to those of GWAS. Considering that differences in environmental distributions across cohorts can reduce the replicability of G×E interactions, and that past G×E studies have suffered from low replication rates, these findings suggest the high reliability of the G×E interactions identified in this manuscript." (**Supplementary Note 4**).

Additionally, we have included the replication in the completely independent cohorts in the **Results** section.

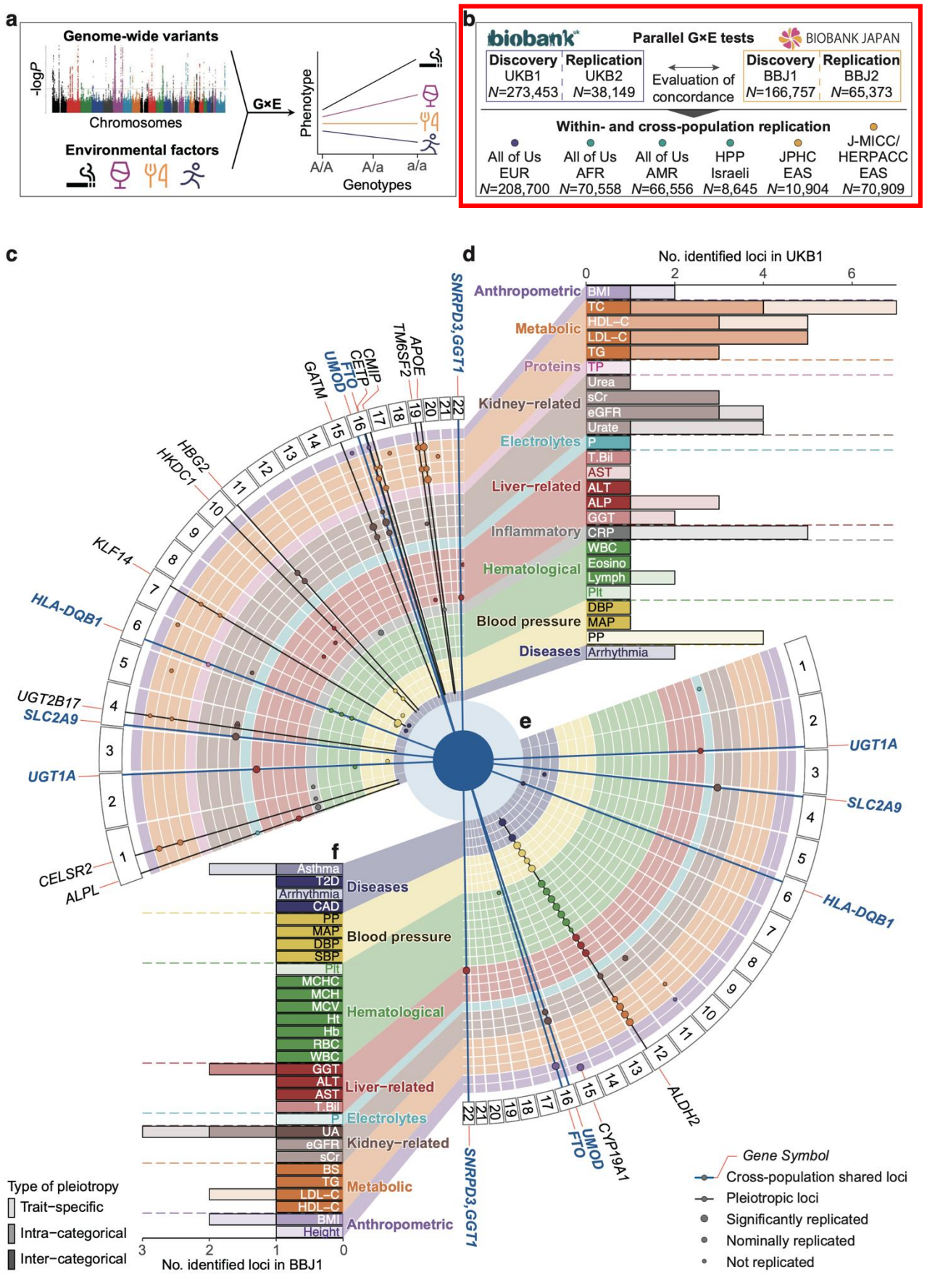
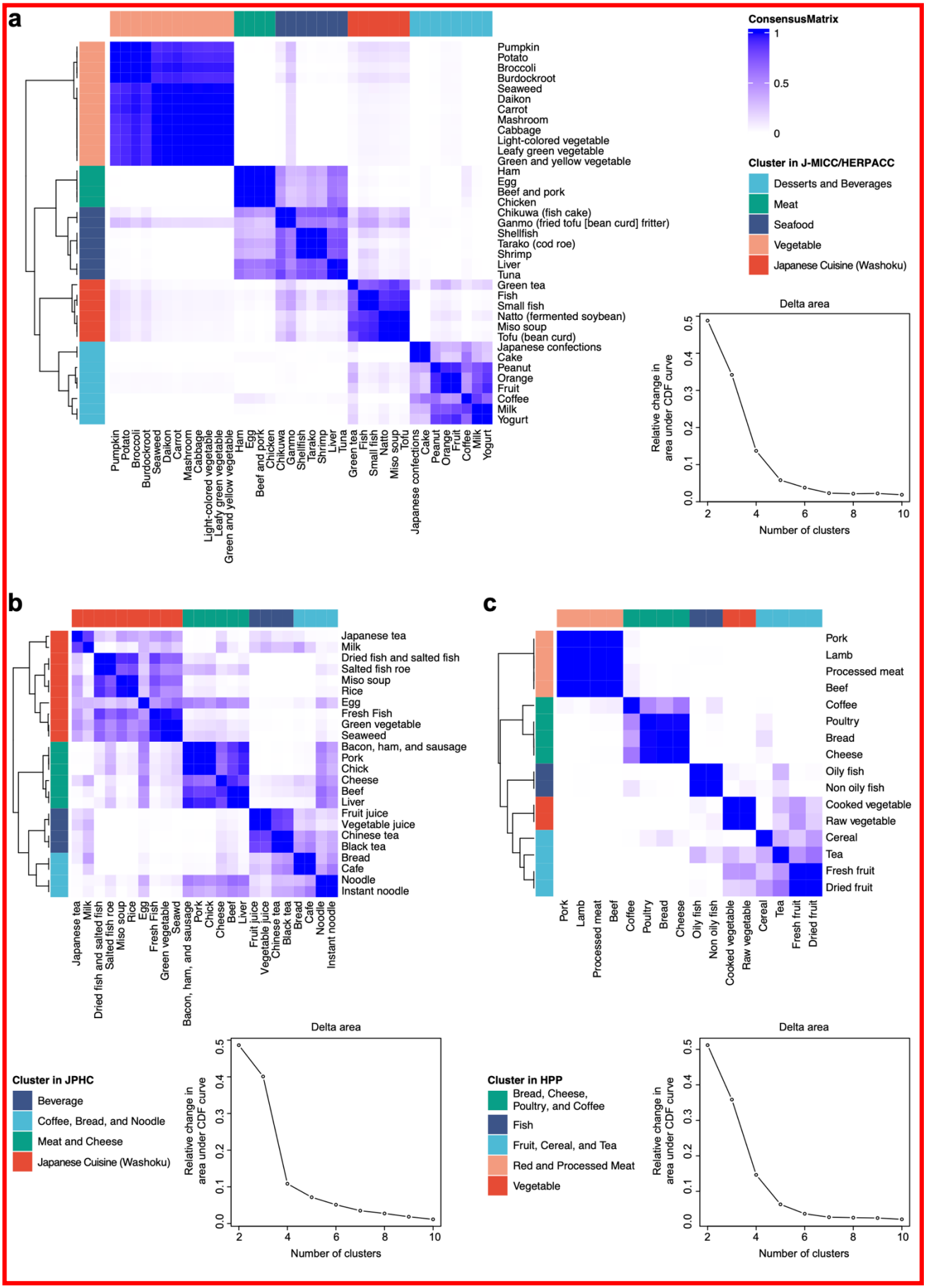

a
Genome-wide variants
-logP
Chromosomes
Environmental factors
G×E
Phenotype
A/A
A/a
a/a
Genotypes
b
biobank
Parallel G×E tests
BIOBANK JAPAN
Discovery
UKB1
N=273,453
Replication
UKB2
N=38,149
Evaluation of concordance
Discovery
BBJ1
N=166,757
Replication
BBJ2
N=65,373
Within- and cross-population replication
All of Us EUR N=208,700
All of Us AFR N=70,558
All of Us AMR N=66,556
HPP Israeli N=8,645
JPHC EAS N=10,904
J-MICC/ HERPACC EAS N=70,909
c
d
No. identified loci in UKB1
Anthropometric
Metabolic
Proteins
Kidney−related
Electrolytes
Liver−related
Inflammatory
Hematological
Blood pressure
Diseases
BMI
TC
HDL−C
LDL−C
TG
TP
Urea
sCr
eGFR
Urate
P
T.Bil
AST
ALT
ALP
GGT
CRP
WBC
Eosino
Lymph
Plt
DBP
MAP
PP
Arrhythmia
SNRPD3,GGT1
APOE
TM6SF2
CMIP
CETP
FTO
UMOD
GATM
HBG2
HKDC1
KLF14
HLA-DQB1
UGT2B17
SLC2A9
UGT1A
CELSR2
ALPL
e
UGT1A
SLC2A9
HLA-DQB1
ALDH2
CYP19A1
FTO
UMOD
SNRPD3,GGT1
f
Asthma
T2D
Arrhythmia
CAD
Diseases
PP
MAP
DBP
SBP
Blood pressure
Plt
MCHC
MCH
MCV
Ht
Hb
RBC
WBC
Hematological
GGT
ALT
AST
T.Bil
Liver−related
P
Electrolytes
UA
eGFR
sCr
Kidney−related
BS
TG
LDL−C
HDL−C
Metabolic
BMI
Height
Anthropometric
No. identified loci in BBJ1
Type of pleiotropy
Trait−specific
Intra−categorical
Inter−categorical
Gene Symbol
Cross-population shared loci
Pleiotropic loci
Significantly replicated
Nominally replicated
Not replicated

Figure 1, A cross-population atlas of G×E interactions

a, a schematic illustration of gene–environment (G×E) interactions. The icons were cited from Phosphor (https://phosphoricons.com/) under MIT license. **b**, an overview of the cohorts used in this study. EUR, AFR, AMR, and EAS denote the European, African, American, and East Asian populations, respectively. **c**, a circular plot of genome-wide significant G×E interactions in UK Biobank (UKB). The dot sizes represent the *P*-values of replication in UKB2 and BBJ2. Pleiotropic loci are shown with the lines and the nearest genes to the lead variants. **d**, the number of genome-wide significant loci for individual traits in UKB. The colors represent the trait categories, and the darkness of the bar colors represents the types of pleiotropy. **e** and **f**, the same as **c** and **d**, respectively, but for Biobank Japan (BBJ). The designs of the figures were inspired by and followed after the referenced studies.

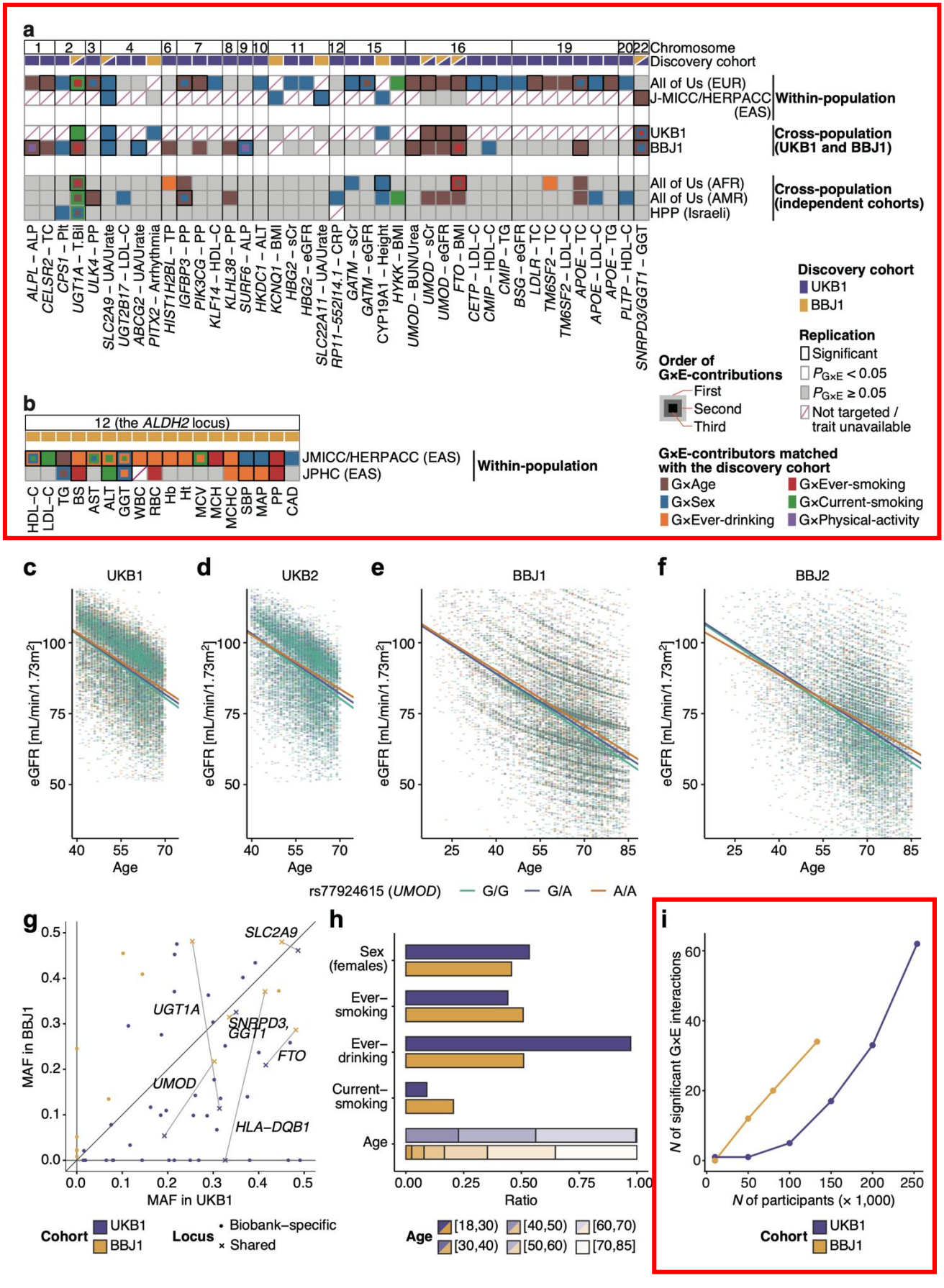
"We further tested replication in independent cohorts: the European population in All of Us (N_{max} = 208,700) and the East Asian population in two Japanese cohorts, J-MICC/HERPACC (N_{max} = 70,909) and JPHC (N_{max} = 10,904), for testing the G×E interactions detected in UKB1 and BBJ1, respectively (**Supplementary Table 10**). Despite the difference in the lifestyle questionnaire between cohorts, clustering of food frequency questionnaires detected dietary clusters consistent to those in the discovery cohorts, including the "Japanese cuisine (Washoku)" cluster observed across all Japanese cohorts and the "meat and cheese" cluster in UKB1 and JPHC, supporting the robustness of our clustering approach (**Ext. Data Fig. 4**). The Bonferroni replication rates was 27% in All of Us (17/62 trait–locus pairs) and 56% in J-MICC/HERPACC (20/36) ($P_{G\times E} < 5.1\times10^{-4}$; **Ext. Data Fig. 5a and b; Supplementary Table 11**). Notably, the pleiotropic G×E interactions at the *ALDH2* locus were replicated at 81% (17/21) in J-MICC/HERPACC, and six of them were again significantly replicated in JPHC, despite its relatively small sample size." (**Results**, P11–12 line 246–258).


a
ConsensusMatrix
1
0.5
0
Pumpkin
Potato
Broccoli
Burdockroot
Seaweed
Daikon
Carrot
Mashroom
Cabbage
Light-colored vegetable
Leafy green vegetable
Green and yellow vegetable
Ham
Egg
Beef and pork
Chicken
Chikuwa (fish cake)
Ganmo (fried tofu [bean curd] fritter)
Shellfish
Tarako (cod roe)
Shrimp
Liver
Tuna
Green tea
Fish
Small fish
Natto (fermented soybean)
Miso soup
Tofu (bean curd)
Japanese confections
Cake
Peanut
Orange
Fruit
Coffee
Milk
Yogurt
Cluster in J-MICC/HERPACC
Desserts and Beverages
Meat
Seafood
Vegetable
Japanese Cuisine (Washoku)
Delta area
Relative change in area under CDF curve
Number of clusters
b
Japanese tea
Milk
Dried fish and salted fish
Salted fish roe
Miso soup
Rice
Egg
Fresh Fish
Green vegetable
Seawed
Bacon, ham, and sausage
Pork
Chick
Cheese
Beef
Liver
Fruit juice
Vegetable juice
Chinese tea
Black tea
Bread
Cafe
Noodle
Instant noodle
Cluster in JPHC
Beverage
Coffee, Bread, and Noodle
Meat and Cheese
Japanese Cuisine (Washoku)
Delta area
Relative change in area under CDF curve
Number of clusters
c
Pork
Lamb
Processed meat
Beef
Coffee
Poultry
Bread
Cheese
Oily fish
Non oily fish
Cooked vegetable
Raw vegetable
Cereal
Tea
Fresh fruit
Dried fruit
Cluster in HPP
Bread, Cheese, Poultry, and Coffee
Fish
Fruit, Cereal, and Tea
Red and Processed Meat
Vegetable
Delta area
Relative change in area under CDF curve
Number of clusters

Extended Data Figure 4. Consensus clustering of dietary questionnaires in independent replication cohorts.

Consensus clustering of food frequency questionnaires in J-MICC/HERPACC (a), JPHC (b), and HPP (c). Delta area plots show the area under the cumulative distribution function (CDF) curve when changing the number of clusters from two to ten.

"We further evaluated replication in the African and American populations in All of Us (N_{max} = 70,558 and 66,556, respectively) and the Israeli population in the Human Phenotype Project (HPP; N_{max} = 8,645). In the African population, three G×E interactions were significantly replicated, two of which overlapped with the shared G×E interactions between UKB1 and BBJ1 (**Ext. Data Fig. 5a and Supplementary Table 13**; $P_{G\times E} < 0.05/77/3 = 2.3\times10^{-4}$, corrected for the number of trait-locus-biobank triplets outside of the *ALDH2* locus and the number of populations). In the American population, two G×E interactions for pulse pressure (PP) were significantly replicated, both originally from UKB1 and primarily driven by G×Age. G×E interactions were not significantly replicated in the Israeli population, possibly due to its relatively small sample size. Nevertheless, we observed the top signal for one of the shared G×E interactions between UKB1 and BBJ1 (*UGT1A* gene family for T.Bil, commonly driven by G×Current-smoking and G×Age)." (**Results**, P13 line 296–307).



Extended Data Figure 5. Replication within and across populations.

a–b, Replication statuses of G×E loci other than the *ALDH2* locus (**a**) and the *ALDH2* locus (**b**). Only interactions with at least one nominally significant replication ($P_{G\times E} < 0.05$) are shown. EUR, the European population; EAS, the East Asian population; AFR, the African population; AMR, the American population. **c–f**, eGFR distributions across rs77924615 genotypes (the *UMOD* locus) and age in UKB1 (**c**), UKB2 (**d**), BBJ1 (**e**), and BBJ2 (**f**), as a representative example of cross-population shared G×E interaction. Dots, 5,000 randomly sampled individuals per genotype; lines, regression lines using all participants. **g and h**, minor allele frequencies (MAF) of the lead variants (**g**) and environmental distributions (**h**) in UKB1 and BBJ1. In **g**, cross-population shared loci are marked with crosses and labeled by nearest genes. Solid gray lines represent $x=0$, $y=0$, and $y=x$. **i**, number of G×E loci detected with $P_{G\times E}$ less than 5.0×10^{-8} in randomly subsampled participants in UKB1 and BBJ1. Subsample sizes: 10, 50, 100, 150, and 200 thousand in UKB1; 10, 50, and 80 thousand in BBJ1. Full-cohort results also shown (N_{mean} = 253,773 in UKB1 and 133,117 in BBJ1).

We described the overview of the replication cohorts in **Supplementary Methods**.

"**Japan Multi-Institutional Collaborative Cohort (J-MICC) Study and Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC)**

J-MICC is a multi-institutional collaboration of prospective and population-based cohorts comprising approximately 100,000 participants in Japan, launched in 2005 with the primary purpose of detecting G×E interactions of lifestyle-related diseases. Individuals aged 35 to 69 years were enrolled from 2005 to 2010 from respondents to study announcements in specified areas, inhabitants attending health checkup examinations by local governments, visitors at health checkup centers, and patients at a cancer hospital. The second assessment was conducted five years after their enrollment. HERPACC is a hospital-based cohort including participants aged 18 or older recruited from 2001 to 2013 from first-visit outpatients at Aichi Cancer Center Hospital. Aichi Cancer Center Hospital is one of the constitutive cohorts of J-MICC, and the same questionnaire was used for both J-MICC and HERPACC. After removing duplicated participants between J-MICC and HERPACC, we analyzed the participants of J-MICC and HERPACC together, which we refer to as J-MICC/HERPACC. For HERPACC, we analyzed the participants with no cancer or history of neoplasm to match the cohort to the population-based characteristics of J-MICC." (**Supplementary Methods**).

"**Japan Public Health Center-based Prospective (JPHC) Study**

The JPHC study is a population-based cohort comprising participants enrolled from residents aged 40-69 years within 11 public health center (PHC) areas nationwide between 1990 and 1994. Before conducting genetic research in the JPHC study, we obtained approval from the institutional review board of the National Cancer Centre (Approval No.: 2011-044), Tokyo, Japan, and provided eligible participants with the option to decline participation. For the present study, 13,024 participants were randomly selected as a subcohort from the JPHC study participants in 9 PHC areas, who had received a health check-up and answered a questionnaire during the baseline survey." (**Supplementary Methods**).

"**All of Us**

All of Us is a nation-wide cohort in the United States, aiming to recruit persons in demographic categories that have been and continue to be underrepresented in biomedical research.

National enrolment began in 2018, and adults 18 years and older who have the capacity to consent and reside in the United States or the United States territory at present are eligible. We analyzed the version 8 dataset of the All of Us." (**Supplementary Methods**).

"**The Human Phenotype Project (HPP)**

HPP is a prospective cohort in Israel, and the participants between the ages of 40 and 70 years were recruited from 2018 to 2021 from the self-assignment of volunteers registered to the trial website. The cohort is largely composed of Ashkenazi Jewish people. Individuals without recent antibiotic use and gastrointestinal morbidity were eligible if they had not yet been diagnosed with the clinical outcomes of interest." (**Supplementary Methods**).

"**Quality control of phenotypes and environments in the replication cohorts**

Clinical traits

We employed the same definition and quality control for clinical traits as those used in UKB1 and BBJ1. In All of Us, biomarker phenotypes were obtained from the electronic medical records, and we used the most recent records for each biomarker. When both medical records from the emergency department and those from other departments were available, we prioritized the latter. We observed frequent mislabeling of the measurement units in this cohort. For example, albumin measurements with the units of mg/mL and mg/L were of the same order of magnitude as those reported in g/dL. Therefore, we corrected the units by examining the median values for individual unit labels. The biomarker phenotypes for the other cohorts were obtained from the baseline assessment data for the main analyses, while we used the second assessment data of J-MICC for the temporal order analyses. We applied the same quality control criteria as used for BBJ and UKB. For disease statuses, we defined participants as cases if they had relevant SNOMED CT codes or self-reported disease histories in All of Us. We used self-reported disease histories for the other cohorts. For JPHC and J-MICC/HERPACC, self-reported histories were recorded for diabetes mellitus instead of T2D, and we used diabetes mellitus as a proxy of T2D for replication purposes.

Environmental factors

The environmental factors from questionnaires, along with the conversion of categorical responses into a continuous scale, were summarized in **Supplementary Table 4**. As only one environmental factor was related to physical activity in JPHC, clustering it with dietary environmental factors would lead to a cluster consisting of both physical activity and dietary factors, opposite to our observation that the "physical activities" cluster was consistently detected as a separate cluster from other dietary clusters in UKB1 and BBJ1. Therefore, we pre-defined the "physical activities" cluster from the environmental factors related to physical activities in the replication cohorts and performed clustering only using dietary factors. In JPHC, two sets of questionnaires with slight differences in questionnaire items were used, and we used the items available in both sets. Questionnaires related to physical activities and dietary consumption were unavailable in All of Us, and we did not perform clustering in this cohort. In All of Us, smoking status was separately asked for cigar smoking, electronic

smoking, hookah smoking, and smokeless tobacco smoking. We defined ever- and current-smoking statuses based on those of all types of tobacco." (**Supplementary Methods**).

4. In relation to #3, I suggest validating the estimated G×E effects by projecting them from the discovery dataset to an independent target dataset. This would help assess whether the estimated G×E interactions have significant predictive value in an external dataset. The use of G×Eprs method, as outlined in https://github.com/DoviniJ/GxEprs <https://github.com/DoviniJ/GxEprs> (Genetic Epidemiology, 48(2), 85-100), could be a useful approach for this purpose.

Response:

Thank you for your thoughtful suggestion and for kindly introducing the G×Eprs method. As described in our response to Comment #3, we first verified the replication of individual G×E loci in completely independent cohorts, including those with different populations and cohort designs. In addition, to validate the G×E effects at the genome-wide level, **we newly constructed G×E-PGS and employed the G×Eprs approach**. As G×E-aware PGS construction methods have not been established, we constructed all types of PGS using PRScs-auto, one of the most widely used tools for PGS construction. **Tested in the newly added independent cohorts (the European population in All of Us and the East Asian population in J-MICC/HERPACC), G×E-PGS significantly explained phenotypic variance for 11/23 and 9/23 trait–environment pairs within and across populations, respectively.** These results provide validation of G×E effects at the genome-wide level and also suggest the potential of G×E-PGS for contributing to phenotype prediction.

Modification:

We described the modification in response to your next comment, #5.

5. The prediction performance of polygenic scores (PGS) is well demonstrated. However, the analyses presented focus more on performance under homogeneous environmental conditions, rather than explicitly modeling G×E interactions. Could the authors conduct similar prediction analyses comparing models with and without G×E interactions? Additionally, do any of the PGS rankings change when incorporating G×E modeling? This comparison would provide valuable insights into the impact of G×E interactions on prediction accuracy.

Response:

Thank you for your thoughtful comment. We performed prediction analysis using G×E-PGS in the revised manuscript. Notably, **the G×E-PGS constructed from G×Sex interactions stratified BMI in opposite directions between males and females in J-MICC/HERPACC** (**Figure 4h**). This stratification was apparent even within each sex-stratified group and among individuals with similar marginal PGS, **suggesting the potential for extending PGS into two dimensions to improve phenotype prediction**. Recent studies have increasingly focused on the interactions between marginal-effect PGS and environments (often referred to as PGS×E) (Boye C *et al.* ***Nat Genet*** (2024)). However, the heterogeneity observed within

individuals of the same sex and with similar marginal-effect PGS cannot be captured by PGS×E, indicating that G×E-PGS more effectively reflects the polygenic architecture underlying sex differences.

A prediction model incorporating G×E-PGS, additive- and marginal-effect PGS, and sex showed a 16% increase in prediction accuracy for BMI compared to the model without G×E-PGS (R^2 = 0.128 versus 0.110, 10-fold cross-validation in J-MICC/HERPACC). These findings demonstrate the utility of G×E-PGS in improving BMI prediction by capturing interaction effects missed by conventional PGS. On the contrary, the gain in prediction accuracy was modest for other trait–environment pairs (**Supplementary Table 25**). As we used biobank-scale large sample sizes and constructed PGS using one of the state-of-the-art methods, these results may show the current challenges in the G×E-PGS framework in general. Larger sample sizes and/or methods tailored for G×E-aware PGS construction are warranted to fully realize the potential of G×E interactions in precision medicine.

Modification:

According to your thoughtful comment, we added the prediction analysis using G×E-PGS.

"Consistent with these findings, PGS constructed from G×E interactions (G×E-PGS) significantly explained phenotypic variance in independent cohorts (11/23 and 9/23 trait–environment pairs within and across populations, respectively; **Supplementary Methods and Supplementary Table 25**). Notably, the G×Sex-based PGS stratified BMI in opposite directions between males and females in J-MICC/HERPACC, capturing the polygenic architecture of sex differences (**Figure 4h**). This stratification was apparent even within each sex and among individuals with similar marginal PGS, suggesting that PGS extended to two dimensions could enhance phenotype prediction. Indeed, a model incorporating G×E-PGS, additive- and marginal-effect PGS, and sex improved BMI prediction accuracy by 16% over a model without G×E-PGS (R^2 = 0.128 versus 0.110, 10-fold cross-validation in J-MICC/HERPACC). In contrast, gains in prediction accuracy were modest for other trait–environment pairs (**Supplementary Table 25**). Larger sample sizes and methods tailored for G×E-PGS construction are warranted to fully realize the potential of G×E interactions in precision medicine." (**Results**, P20–21 line 485–498).

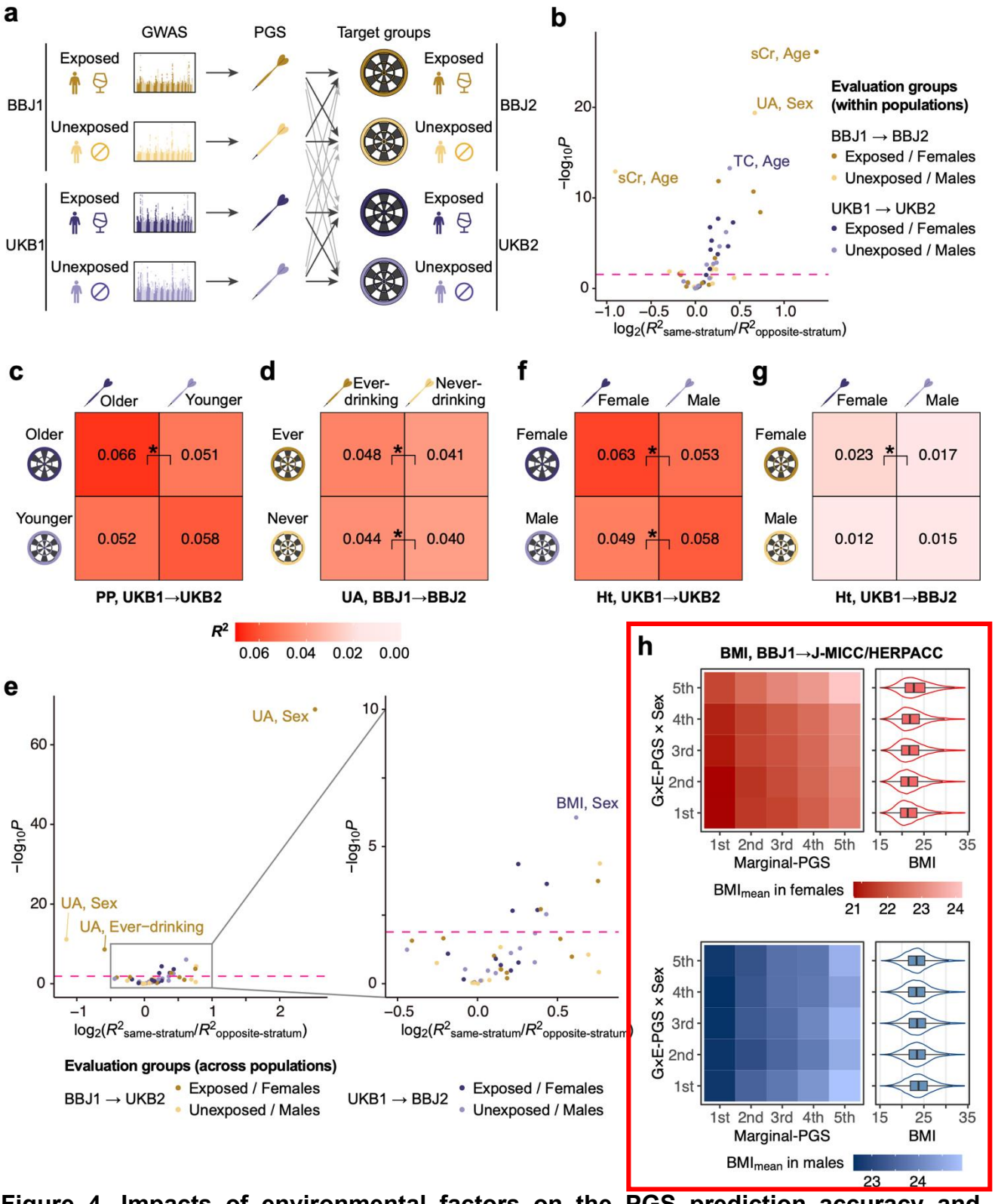


Figure 4, Impacts of environmental factors on the PGS prediction accuracy and portability

a, overview of polygenic score (PGS) construction and evaluation of within-population prediction accuracy and cross-population portability. Wine icons from Phosphor (https://phosphoricons.com/) under MIT license. **b**, PGS prediction accuracy in the same population. *X*-axis: $\log_2$ ratio of prediction accuracy (R^2) between "same-stratum" and "opposite-stratum" PGS (based on matching vs. non-matching environmental strata). *Y*-axis:

P-values of R^2 differences. Dashed line: FDR threshold of 0.05. **c** and **d**, examples of within-population prediction accuracy (R^2): pulse pressure (PP) in UKB stratified by age (**c**) and uric acid (UA) in BBJ stratified by drinking status (**d**). Asterisks: FDR < 0.05. Figure designs inspired by the referenced studies. **e**, same as **b**, for cross-population portability. **f** and **g**, examples of PGS prediction accuracy within (**f**) and across (**g**) populations, for hematocrit (Ht) stratified by sex. **h**, BMI distribution stratified by PGS constructed from G×E interactions (G×E-PGS) and that from marginal PGS. Boxplots show median, quartiles, and 1.5× interquartile range whiskers.

"**Phenotype prediction using PGS constructed from G×E interactions (G×E-PGS)**

The G×Eprs framework

We employed the G×Eprs framework to test whether G×E-PGS significantly explains the phenotypic variation. This framework fits the following linear regression model:

$$Y \sim PGS_{additive} + PGS_{G\times E} \times E + PGS_{G\times E} + E$$

where $PGS_{additive}$ is the PGS constructed from additive genetic effects, $PGS_{G\times E}$ is PGS constructed from G×E interaction effects, and E denotes the environmental factor. We also refer to the PGS constructed from marginal effects as $PGS_{marginal}$ below. As G×E-aware PGS construction methods have not been established, we constructed all types of PGS using PRScs-auto with the 1000 Genome Project reference panel for the corresponding population. For the response variable Y, we used phenotype residuals obtained by regressing the raw phenotype values on the same covariates used in the G×E interaction tests. To preserve the variance attributable to environmental factors, we excluded age×sex, age^2, and age^2×sex from the covariates when testing G×Age PGS and excluded age×sex and age^2×sex when testing G×Sex PGS. Following the original study, we evaluated the significance of the $PGS_{G\times E} \times E$ term using both Wald *P*-values and permutation-based *P*-values, and considered it significant when FDR was less than 0.05 for both *P*-values. Permutation tests were performed with 10,000 iterations by randomly shuffling the values of $PGS_{G\times E} \times E$, as implemented in the R package "GxEprs".

Phenotype prediction using G×E-PGS

We compared two elastic net models for phenotype prediction, one with G×E-PGS and one without G×E-PGS. The baseline model included marginal PGS and the environmental factor (E), assuming a standard PGS constructed based on GWAS summary statistics:

$$Y \sim PGS_{marginal} + E$$

The full model additionally included additive PGS, G×E-PGS, and the interaction with the environment:

$$Y \sim PGS_{marginal} + PGS_{additive} + PGS_{G\times E} + PGS_{G\times E} \times E + E$$

We also included Age^2 into both models when testing G×Age-PGS to account for non-linear variation in phenotype values across age. Models were implemented using the R package "glmnet" (version 4.1.8), which performs regularized regression by linearly combining L1 (Lasso) and L2 (Ridge) penalties. The hyperparameters (Lasso and Ridge penalties) were tuned via a grid search over a range from 0 to 1 in increments of 0.1, and were selected based on 10-fold cross-validation by minimizing the mean squared error on the validation

folds. We reported the R^2 from the 10-fold cross-validation as the prediction accuracy." (**Supplementary Methods**).

6. The omic-wide interaction models are not thoroughly described in the Methods section. The authors should provide more detailed information about the model, including its implementation and underlying assumptions. This is particularly important because non-genomic omics data are more susceptible to being influenced by outcomes compared to genomic variables. A precise and comprehensive description of the model would help address potential concerns related to reverse causality and ensure clarity for reproducibility.

Response:

Thank you for your valuable comment. We agree that a precise and comprehensive description of the statistical models is essential to ensure reproducibility and to address concerns such as potential reverse causality. We used omics data as phenotypes and performed G×E interaction tests using the same analytical framework as in the main analyses of disease and biomarker traits. That is, **we tested for interactions between genotype and environmental exposures (G×E) on omics traits, rather than omics-by-environment interactions**. As described in our response to Comment #3, the only difference in model specification was the inclusion of statin usage as an additional covariate for the metabolome, due to its well-known influence on metabolite levels.

In addition, while we initially tested G×E interactions for omics data only in the lead variants of the main analyses, **we have newly conducted genome-wide G×E interaction tests for the metabolome in the revised manuscript**, which benefited from its large sample size. **We increased the sample sizes in BBJ1 from 43,371 to 89,040**, substantially enhancing the power of the interaction analyses. Conducting genome-wide tests would help clarify that our analyses focused on G×E interactions, rather than omics-by-environment interactions.

We identified 30 and 11 genome-wide significant loci in UKB1 and BBJ1, respectively, yielding 736 and 228 metabolite–locus pairs. Among these, 16 loci in UKB1 and 4 in BBJ1 passed the Bonferroni-corrected threshold ($P_{G\times E} < 7.7\times10^{-11}$, correcting for 325 metabolites across two cohorts). Notably, G×Sex interactions at the *ALDH1A2* and *ZNF259* loci were detected in both cohorts. In addition, several loci exhibited sex-dimorphic G×Sex interactions, with opposite directions of association between males and females, including the *LPL*, *FADS2/FADS1*, *APOB*, and *GLDC* loci, further underscoring the importance of sex-specific genetic architecture in metabolome regulation. Although not reaching the stringent Bonferroni threshold, we also reported notable G×E interactions, included a current-smoker-specific association at the *LRRC4C* locus for phospholipids to total lipids ratio in large VLDL (L_VLDL_PL_pct, $P_{G\times E} = 1.1\times10^{-8}$), a never-drinker-specific association at the *KLHL32* locus for phospholipids in small VLDL (S_VLDL_PL, $P_{G\times E} = 4.8\times10^{-9}$), and a G×E interaction without a marginal effect for acetone ($P_{G\times E} = 2.2\times10^{-8}$) at the *SLC7A2* locus, driven by tea and coffee consumption, ever smoking, and meat and cheese consumption.

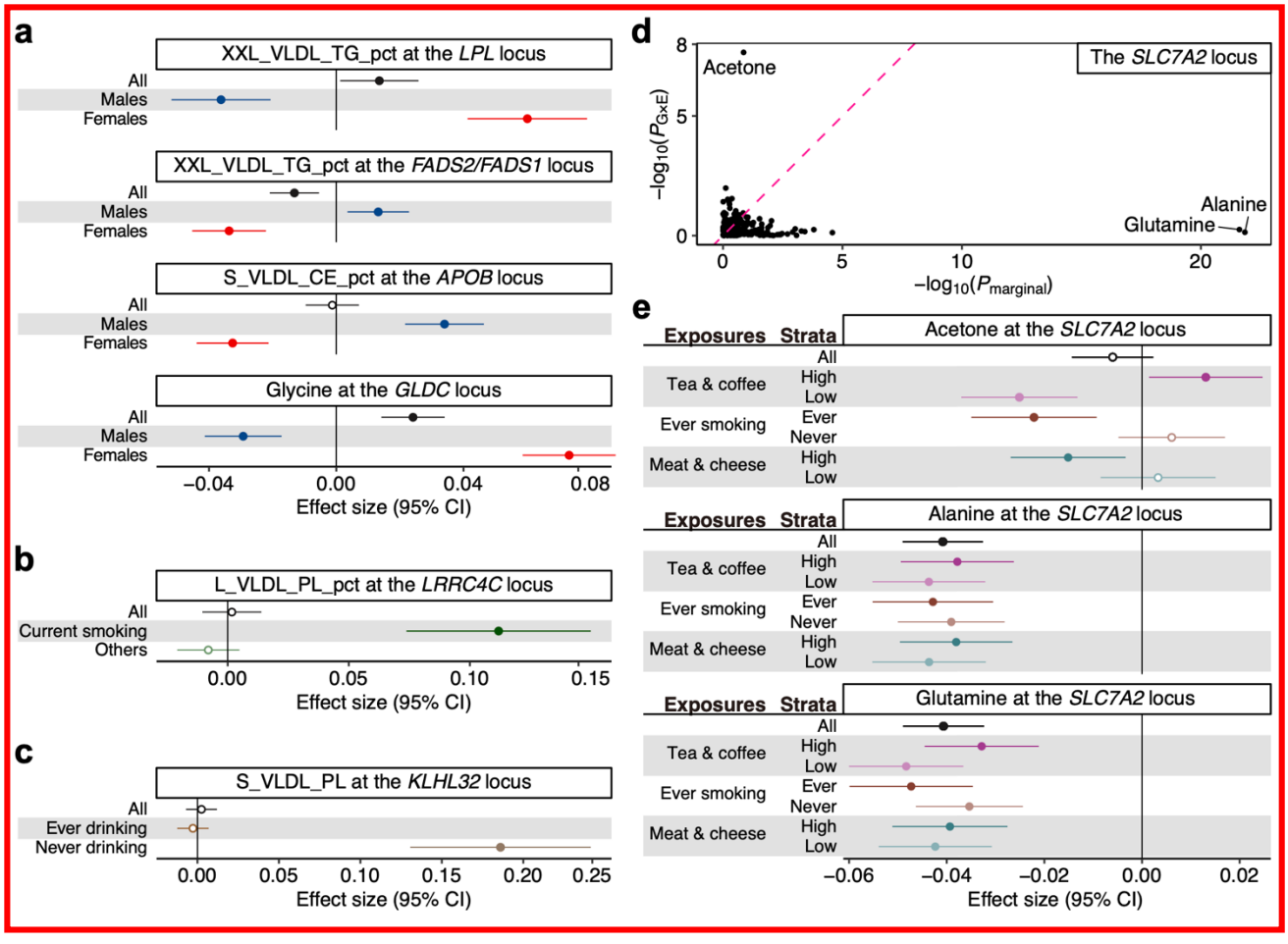
Modification:

We have revised the **Methods** to describe our methods more clearly, as written in Comment #3. In addition, we included genome-wide G×E interaction tests for the metabolome in the **Results** section.

"Motivated by these G×E metabolite QTLs, we expanded the metabolome-wide G×E analysis to genome-wide variants, identifying 30 and 11 genome-wide significant loci in UKB1 and BBJ1, respectively, yielding 736 and 228 metabolite–locus pairs. Among these, 16 loci in UKB1 and 4 in BBJ1 passed the Bonferroni threshold ($P_{G\times E} < 7.7\times10^{-11}$, correcting for 325 metabolites across two cohorts). Notably, G×Sex interactions at the *ALDH1A2* and *ZNF259* loci were observed in both cohorts (**Supplementary Table 31**). Several loci exhibited sex-dimorphic G×Sex interactions with opposite effect directions in males and females. These included the *LPL*, *FADS2/FADS1*, *APOB*, and *GLDC* loci (**Ext. Data Fig. 10**; $P_{G\times E} = 8.7\times10^{-24}$, 7.7×10^{-27}, 6.6×10^{-17}, and 4.9×10^{-28}, respectively), again underscoring the sex-specific genetic architecture of metabolome regulation. Additional notable G×E signals, though below Bonferroni significance, included a current-smoker-specific association at the *LRRC4C* locus for phospholipids to total lipids ratio in large VLDL (L_VLDL_PL_pct; $P_{G\times E} = 1.1\times10^{-8}$), and a never-drinker-specific association at the *KLHL32* locus for phospholipids in small VLDL (S_VLDL_PL; $P_{G\times E} = 4.8\times10^{-9}$). At the *SLC7A2* locus, we observed marginal effects for alanine and glutamine and a G×E-only effect for acetone ($P_{G\times E} = 2.2\times10^{-8}$), driven by tea and coffee consumption, ever smoking, and meat and cheese consumption (**Ext. Data Fig. 10**). Further validation and functional studies are warranted to elucidate their underlying mechanisms." (**Results**, P25–26 line 615–632).

"We further conducted omics-level G×E QTL analyses to pinpoint the metabolites and proteins that mediate clinical G×E interactions. Multiple sex-dimorphic and environment-specific genetic effects on the metabolome suggest that omics data may reveal G×E interactions with large effect sizes." (**Discussion**, P27–28 line 660–665).

"For the genome-wide G×E interaction tests for metabolome, we targeted (i) genotyped variants in each cohort, (ii) HapMap3 variants, and (iii) the variants targeted by the meta-analysis across BBJ and UKB, to reduce the computational cost. After defining the metabolome G×E loci based on their distance as above, we merged the metabolome G×E loci if they overlapped with the same locus defined for the clinical traits." (**Methods**, P44–45 line 1124–1129).



Extended Data Figure 10. Genome-wide G×E interactions for metabolites.

a, sexual dimorphic effects at four loci in UKB1. XXL_VLDL_TG_pct, the triglycerides to total lipids ratio in chylomicrons and extremely large VLDL; S_VLDL_CE_pct, the cholesteryl esters to total lipids ratio in small VLDL. **b and c**, environment-specific effects in UKB1 for current-smoking at the *LRRC4C* locus (**b**) and never-drinking at the *KLHL32* locus (**c**). L_VLDL_PL_pct, the phospholipids to total lipids ratio in large VLDL; S_VLDL_PL, phospholipids in small VLDL. **d**, distribution of $P_{marginal}$ and $P_{G\times E}$ at the *SLC7A2* locus in UKB1. **e**, environment-stratified effect sizes at the *SLC7A2* locus in UKB1. For the "tea and coffee" and "meat and cheese" environment clusters, participants were stratified into two equal-sized groups. Note that coffee consumption was multiplied by -1 to account for its negative correlation with tea consumption in UKB1; therefore, higher cluster values indicate more frequent tea consumption, whereas lower values indicate more frequent coffee consumption.

Reviewers #2 and #3:

To the author

This article by Namba et. al. demonstrates an excellent application of detecting gene environment interactions in two cohorts. It expands on current knowledge by providing an estimate of cross population replication, estimating missing heritability attributed to GxE, and taking a closer examination on the biology behind top GxEs. The authors also utilize single-cell transcriptomics, metabolomics, and proteomics data to connect interactions to molecular markers. Strengths include methodology, clear writing, and beautiful visualizations. The analyses and results are not as comprehensive or novel as claimed, and this is my largest concern for publication in Nature.

Response:

We truly thank you for carefully reviewing our manuscript and kindly acknowledging the value of our study, particularly our in-depth analyses of the biology behind G×E interactions. We appreciate your many thoughtful comments, which substantially contributed to further improvement of the manuscript. Please find our responses and additional analyses below for the comprehensiveness (#1), novelty and multiple testing correction (#2–6), and other comments. To improve readability, the changes made to the figures are highlighted with red boxes.

Comments

1. The authors greatly over-state the impact of their findings with descriptors such as “compendium”, “atlas”, “[phenome-, environment-, etc] -wide”, “comprehensive”, “catalog”, and “cross-population”. They only investigated 47 phenotypes, 12 environments (which also include less “environmental” exposures such as age and sex), and 2 biobanks of 2 different ancestries, making a much more focused analysis, and do not provide a “catalog” or “atlas” resource for readers to interrogate a “comprehensive” set of GxEs.

Response:

Thank you for your comment. We appreciate your point regarding the terminology used to describe the scope of our study. Our intention was to emphasize the broad and integrative nature of our analysis, which we believe is the best available at the current stage, compared to previous genome-wide G×E studies. Prior studies have focused on a single category of phenotypes or environments, often with a limited number of phenotypes or environments, or have targeted specific loci rather than conducting genome-wide analyses. For example, Moore R *et al.* ***Nat Genet*** (2019) utilized UKB to investigate G×E interactions for one phenotype, BMI, and examined 97 variants. Bernabeu E *et al.* ***Nat Genet*** (2021) has tested G×E in a phenome-wide manner in UKB, but they focused only on G×Sex interactions. Westerman KE *et al*. ***Nat Commun*** (2022) investigated G×E in an environment-wide manner using UKB, focusing on serum cardiometabolic biomarkers, and did not examine G×E for genome-wide variants. Therefore, the global overview of G×E interactions remained unclear,

including the genome-wide G×E interaction mapping, how multiple environments drove G×E interactions at individual loci, and how multiple phenotypes were affected by the same G×E interactions (i.e., pleiotropy). **Our study systematically tested all possible combinations of 47 phenotypes and 12 environmental factors (9 factors for each cohort), resulting in a total of 846 phenotype–environment–population triplets, using genome-wide variants**. Additionally, previous reports have primarily focused on the European population, mainly using UKB. We are aware of a few exceptions conducted by the CHARGE consortium (Bentley AR *et al.* ***Nat Genet*** (2019), Laville V *et al.* ***EJHG*** (2022)), while they focused on serum lipids and blood pressure for phenotypes and drinking and smoking for environmental factors. Therefore, we investigated G×E interactions in two populations in parallel to depict the cross-population consistency of G×E interactions for a wide range of phenotype–environment pairs.

Unlike GWAS, G×E analyses require sufficient sample sizes within subsets stratified by environmental factors. In line with a previous review that discussed the need for substantially larger sample sizes for G×E interaction tests compared to GWAS (Aschard H. ***Genet Epidemiol*** (2016)), **we newly conducted a subsampling analysis of UKB1 and BBJ1, revealing that biobank-scale sample sizes were required to detect G×E interactions** (**Ext. Data Fig. 5i**). To ensure sufficient statistical power for genome-wide analyses, we focused on 47 clinical traits available in both UKB1 and BBJ1. These included 38 quantitative biomarkers and 9 binary diseases with a case–control ratio greater than 1:15 in at least one biobank, spanning 10 categories to enable an evaluation of pleiotropy. Notably, we found that pleiotropy of G×E interactions was trait-category specific, in contrast to the frequent inter-categorical pleiotropy observed in marginal effects. The threshold of the case–control ratio was chosen to mitigate test statistic inflation caused by case–control imbalance in logistic regression. For dichotomous traits, we re-estimated effect sizes and *P*-values using Firth logistic regression for the variants with *P*-values less than 5.0×10^{-8}. As Firth logistic regression is computationally intensive, applying it to traits with more extreme case–control imbalances would have generated numerous spurious signals and rendered the analysis computationally infeasible. Therefore, instead of performing genome-wide analyses for these unbalanced traits, we conducted PheWAS for 35 diseases in UKB1 and 52 diseases in BBJ1.

To perform G×E interaction tests in UKB and BBJ in parallel, we used the environments available in both biobanks. We employed a clustering approach to group dietary and physical activity questionnaires, following the strategy in Westerman KE *et al.* ***Hum Mol Genet*** (2021), with the aim of enhancing the detection power of G×E interactions and reducing multicollinearity among highly correlated questionnaire items. We included 21 and 15 raw questionnaires for the clustering analysis in UKB1 and BBJ1, respectively, and summarized them into four clusters in both biobanks. This approach was particularly helpful in identifying the environmental factor that contributed to the G×E interaction at the *ABCG2* locus, where meat consumption consistently drove the G×E interaction across meat types. This result demonstrates that our clustering approach can provide a better interpretation of the G×E interaction than analyzing individual food items separately. Therefore, we prioritized using environmental clusters, instead of unnecessarily increasing the number of environmental variables by including raw questionnaire items.

We included age and sex in our analysis as they represent biologically and socially meaningful sources of environmental variation. Age captures both senescence and age-related social environments, and we demonstrated the biological relevance of G×Age interactions in our single-cell analysis, which revealed age-dependent shifts in the cell types responsible for pulse pressure. Sex reflects both the hormone milieu and differential environmental pressures arising from social gender roles, as discussed in Bernabeu E *et al.*

Nat Genet (2021), which you kindly referenced in Comment #6. Importantly, G×E interaction analyses allowed us to systematically and simultaneously model diverse sources of environmental variation, including age, sex, and other lifestyle factors, within a unified analytical framework. This underscores the broad applicability of the G×E approach.

To make our findings accessible, **we will publicly release full summary statistics of all genome-wide G×E interaction tests**, including those for all possible phenotype–environment pairs, providing a valuable resource to the research community. We have shown that this dataset enabled not only replication and secondary analysis but also offered a means to study phenomena such as G×E pleiotropy and G×E interactions driven by multiple environments, as exemplified by the highly pleiotropic *ALDH2* locus in BBJ and the *FTO* locus driven by multiple environments in both UKB and BBJ. Our systematic analyses also enabled us to evaluate the genome-wide property of G×E interactions, including heritability and cross-trait correlations, contributing to depict the global overview of G×E interactions. Importantly, these analyses could not be done without targeting wide ranges of phenotypes and environments and genome-wide variants.

Additionally, we performed the G×E interaction tests for omics data (325 metabolites and 2,924 proteins). While our initial manuscript limited the omics-wide G×E test to the lead variants, **we have now conducted genome-wide G×E interaction tests for metabolites in both UKB1 and BBJ1, leveraging their large sample sizes. We increased the sample sizes in BBJ1 from 43,371 to 89,040**, substantially enhancing the power of the interaction analyses. We identified 30 and 11 genome-wide significant loci in UKB1 and BBJ1, respectively, yielding 736 and 228 metabolite–locus pairs. Among these, 16 loci in UKB1 and 4 in BBJ1 passed the Bonferroni-corrected threshold ($P_{G\times E} < 7.7\times10^{-11}$, correcting for 325 metabolites across two cohorts). Notably, G×Sex interactions at the *ALDH1A2* and *ZNF259* loci were detected in both cohorts. In addition, several loci exhibited sex-dimorphic G×Sex interactions, with opposite directions of association between males and females, including the *LPL*, *FADS2/FADS1*, *APOB*, and *GLDC* loci, further underscoring the importance of sex-specific genetic architecture in metabolome regulation. Although not reaching the stringent Bonferroni threshold, we also reported notable G×E interactions, included a current-smoker-specific association at the *LRRC4C* locus for phospholipids to total lipids ratio in large VLDL (L_VLDL_PL_pct, $P_{G\times E} = 1.1\times10^{-8}$), a never-drinker-specific association at the *KLHL32* locus for phospholipids in small VLDL (S_VLDL_PL, $P_{G\times E} = 4.8\times10^{-9}$), and a G×E interaction without a marginal effect for acetone ($P_{G\times E} = 2.2\times10^{-8}$) at the *SLC7A2* locus, driven by tea and coffee consumption, ever smoking, and meat and cheese consumption. Our systematic genome-wide G×E interaction tests for the metabolome, which covered 325 × 9 × 2 = 5,850 phenotype–environment–population triplets, further enhanced the resource value of our manuscript.

Regarding populations, we completely agree with you about the importance of expanding population diversity. Previous studies have primarily focused on European populations, and the cross-population consistency of G×E interactions has not been systematically examined for any non-European population. To the best of our knowledge, the cross-population consistency was systematically evaluated in our manuscript for the first time. As discussed above, the subsampling of UKB1 and BBJ1 revealed that discovering G×E interactions required biobank-scale large sample sizes. This result suggests that focusing on the European and East Asian populations at the discovery stage was reasonable, considering the current availability of the data with genomes, phenotypes, and environments.

To further enhance population diversity, we tested for replication in the newly added four independent cohorts with diverse populations (All of Us, J-MICC/HERPACC,

JPHC, and HPP; comprising European, East Asian, African, American, and Israeli populations; *N* = 436,272). Before evaluating replication in a cross-population manner, we first assessed within-population replication of the G×E interactions in completely independent cohorts. To this end, we employed the European population in All of Us (N_{max} = 208,700; in the US) and the East Asian population in J-MICC/HERPACC (N_{max} = 70,909; in Japan) for replicating the UKB1 and BBJ1 results, respectively. In these cohorts, 27% (17/62) and 56% (20/36) trait–locus pairs were significantly replicated ($P_{G\times E} < 0.05/(62+36) = 5.1\times10^{-4}$), respectively. Notably, the pleiotropic G×E interactions at the *ALDH2* locus showed a particularly high replication rate in J-MICC/HERPACC (81% = 17/21). We also tested replication in another Japanese cohort, JPHC (N_{max} = 10,904), where 6 G×E interactions at the *ALDH2* locus were again significantly replicated.

Subsequently, we evaluated the cross-population replication of the G×E interactions in the African and American populations from All of Us (N_{max} = 70,558 and 66,556, respectively) and the Israeli population from HPP (N_{max} = 8,645, largely composed of Ashkenazi Jewish people). In the African population, three G×E interactions were significantly replicated, and two of them were included in the shared G×E interactions between UKB1 and BBJ1. Two G×E interactions for pulse pressure were significantly replicated in the American population, both of which were originally detected in UKB1 and primarily driven by G×Age in both UKB1 and All of Us. G×E interactions were not significantly replicated in the Israeli population, possibly due to its relatively small sample size. Nevertheless, we observed the top signal for one of the shared G×E interactions between UKB1 and BBJ1 (*UGT1A* gene family for total bilirubin, commonly driven by G×Current-smoking and G×Age), suggesting its consistency across diverse populations. We summarized the replication statuses, including nominal replication with $P_{G\times E}$ less than 0.05, in **Ext. Data Fig. 5a and b**. We note that we stringently verified that all replicated G×E interactions were driven by the same environments as the discovery cohorts, and the effect sizes were in the same direction. We regarded any G×E interactions that did not meet these criteria as non-replicated, regardless of their *P*-values.

Among the independent cohorts, All of Us is a nationwide cohort, and J-MICC/HERPACC and JPHC are population-based cohorts; the participants of HPP were recruited from volunteers in Israel. **As these cohorts represent diverse populations with different recruitment strategies, questionnaires, and healthcare systems, replication in them supports the generalizability of our findings across heterogeneous settings**.

In summary, our manuscript reported genome-wide G×E interactions for a broad range of phenotypes, environments, omics layers, and populations. We tested G×E interactions for all possible combinations of phenotypes and environments, and our additional genome-wide analyses of the metabolome substantially increased the number of phenotype–environment pairs. We will make their full summary statistics publicly available upon acceptance. Following your comment, we tested replication in five diverse populations, with a total sample size of 980,004 from six cohorts across the discovery and replication stages. We believe that our manuscript represents the most comprehensive effort to date in characterizing the landscape of G×E interactions at scale.

Modification:

We added the replication tests in completely independent cohorts with diverse populations in the **Results** section.

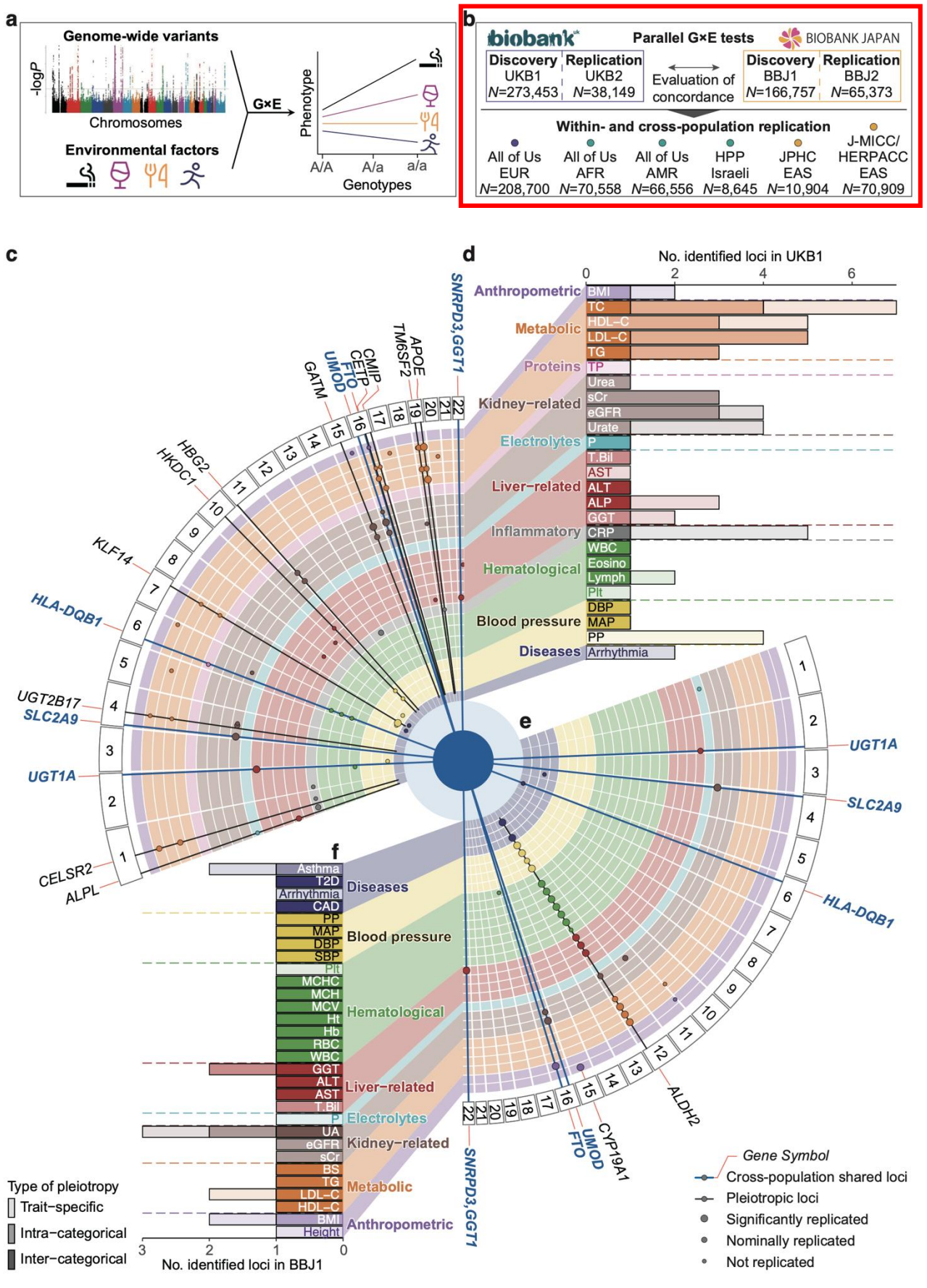
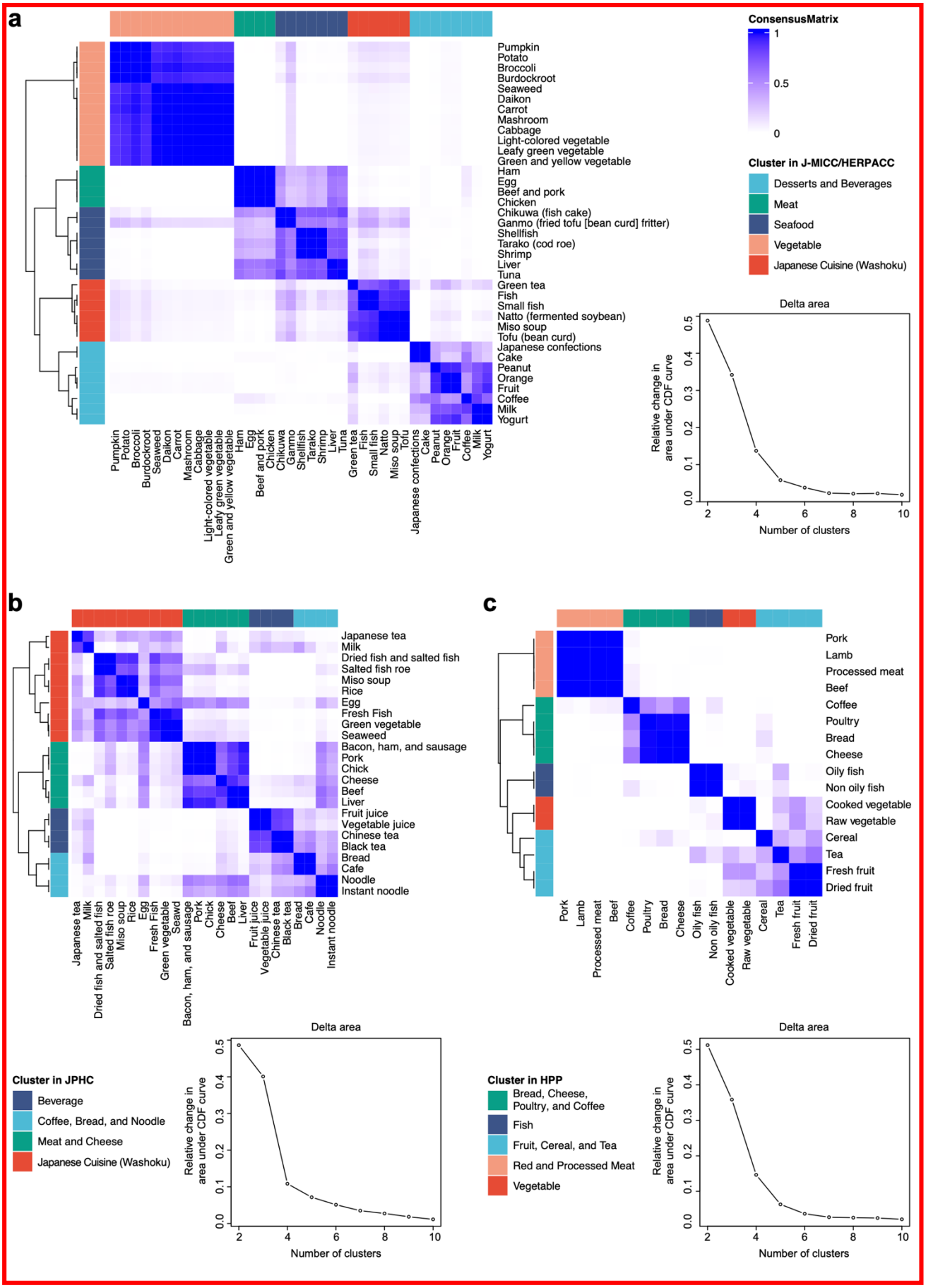

a
Genome-wide variants
-logP
Chromosomes
Environmental factors
G×E
Phenotype
A/A
A/a
a/a
Genotypes
b
biobank
Parallel G×E tests
BIOBANK JAPAN
Discovery
UKB1
N=273,453
Replication
UKB2
N=38,149
Evaluation of concordance
Discovery
BBJ1
N=166,757
Replication
BBJ2
N=65,373
Within- and cross-population replication
All of Us EUR N=208,700
All of Us AFR N=70,558
All of Us AMR N=66,556
HPP Israeli N=8,645
JPHC EAS N=10,904
J-MICC/ HERPACC EAS N=70,909
c
d
No. identified loci in UKB1
Anthropometric
Metabolic
Proteins
Kidney−related
Electrolytes
Liver−related
Inflammatory
Hematological
Blood pressure
Diseases
BMI
TC
HDL−C
LDL−C
TG
TP
Urea
sCr
eGFR
Urate
P
T.Bil
AST
ALT
ALP
GGT
CRP
WBC
Eosino
Lymph
Plt
DBP
MAP
PP
Arrhythmia
SNRPD3,GGT1
APOE
TM6SF2
CMIP
CETP
FTO
UMOD
GATM
HBG2
HKDC1
KLF14
HLA-DQB1
UGT2B17
SLC2A9
UGT1A
CELSR2
ALPL
e
UGT1A
SLC2A9
HLA-DQB1
ALDH2
CYP19A1
FTO
UMOD
SNRPD3,GGT1
f
Asthma
T2D
Arrhythmia
CAD
Diseases
PP
MAP
DBP
SBP
Blood pressure
Plt
MCHC
MCH
MCV
Ht
Hb
RBC
WBC
Hematological
GGT
ALT
AST
T.Bil
Liver−related
P
Electrolytes
UA
eGFR
sCr
Kidney−related
BS
TG
LDL−C
HDL−C
Metabolic
BMI
Height
Anthropometric
No. identified loci in BBJ1
Type of pleiotropy
Trait−specific
Intra−categorical
Inter−categorical
Gene Symbol
Cross-population shared loci
Pleiotropic loci
Significantly replicated
Nominally replicated
Not replicated

Figure 1, A cross-population atlas of G×E interactions

a, a schematic illustration of gene–environment (G×E) interactions. The icons were cited from Phosphor (https://phosphoricons.com/) under MIT license. **b**, an overview of the cohorts used in this study. EUR, AFR, AMR, and EAS denote the European, African, American, and East Asian populations, respectively. **c**, a circular plot of genome-wide significant G×E interactions in UK Biobank (UKB). The dot sizes represent the *P*-values of replication in UKB2 and BBJ2. Pleiotropic loci are shown with the lines and the nearest genes to the lead variants. **d**, the number of genome-wide significant loci for individual traits in UKB. The colors represent the trait categories, and the darkness of the bar colors represents the types of pleiotropy. **e** and **f**, the same as **c** and **d**, respectively, but for Biobank Japan (BBJ). The designs of the figures were inspired by and followed after the referenced studies.

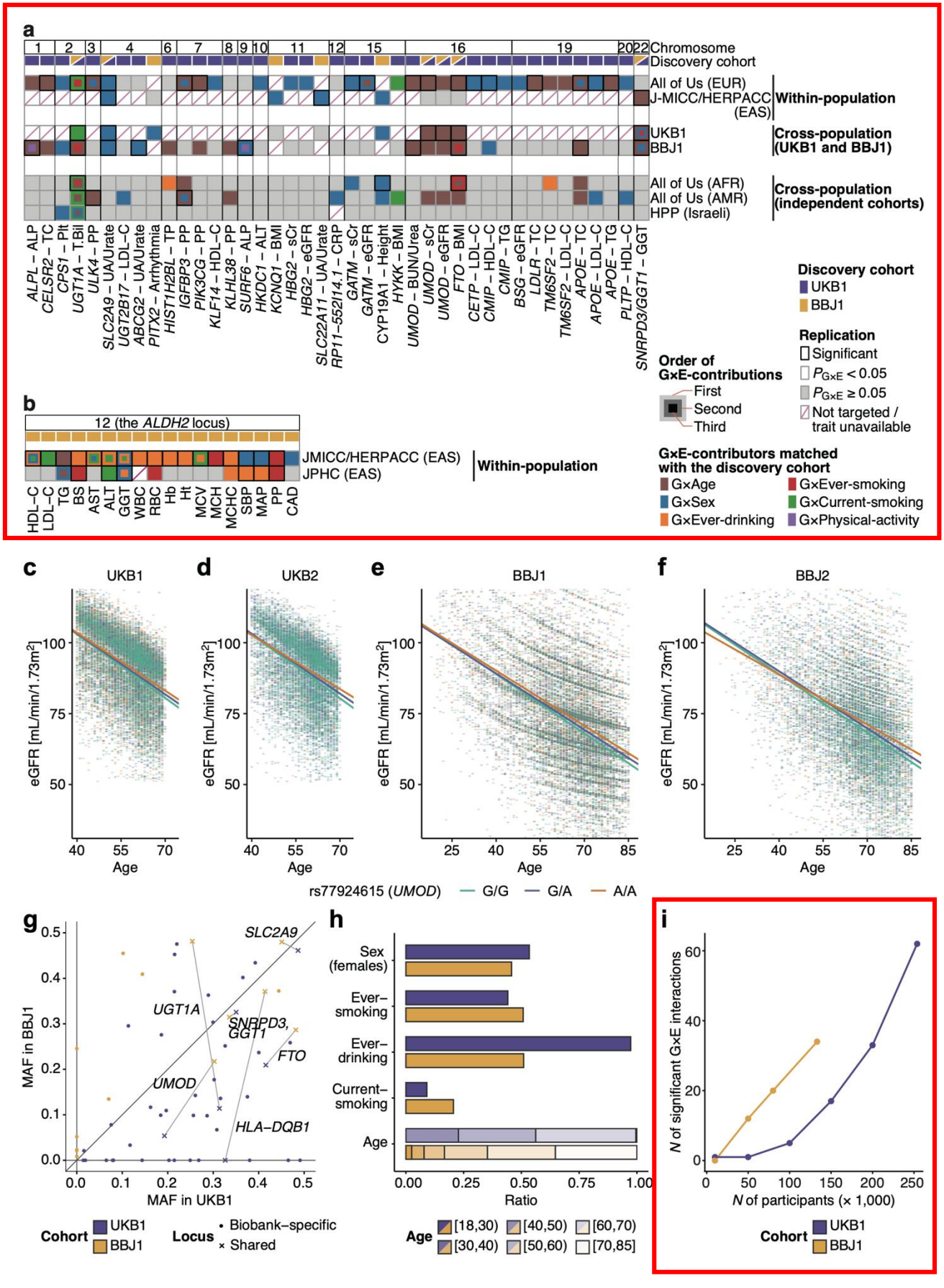
"We further tested replication in independent cohorts: the European population in All of Us (N_{max} = 208,700) and the East Asian population in two Japanese cohorts, J-MICC/HERPACC (N_{max} = 70,909) and JPHC (N_{max} = 10,904), for testing the G×E interactions detected in UKB1 and BBJ1, respectively (**Supplementary Table 10**). Despite the difference in the lifestyle questionnaire between cohorts, clustering of food frequency questionnaires detected dietary clusters consistent to those in the discovery cohorts, including the "Japanese cuisine (Washoku)" cluster observed across all Japanese cohorts and the "meat and cheese" cluster in UKB1 and JPHC, supporting the robustness of our clustering approach (**Ext. Data Fig. 4**). The Bonferroni replication rates was 27% in All of Us (17/62 trait–locus pairs) and 56% in J-MICC/HERPACC (20/36) ($P_{G\times E} < 5.1\times10^{-4}$; **Ext. Data Fig. 5a and b; Supplementary Table 11**). Notably, the pleiotropic G×E interactions at the *ALDH2* locus were replicated at 81% (17/21) in J-MICC/HERPACC, and six of them were again significantly replicated in JPHC, despite its relatively small sample size." (**Results**, P11–12 line 246–258).


a
ConsensusMatrix
1
0.5
0
Pumpkin
Potato
Broccoli
Burdockroot
Seaweed
Daikon
Carrot
Mashroom
Cabbage
Light-colored vegetable
Leafy green vegetable
Green and yellow vegetable
Ham
Egg
Beef and pork
Chicken
Chikuwa (fish cake)
Ganmo (fried tofu [bean curd] fritter)
Shellfish
Tarako (cod roe)
Shrimp
Liver
Tuna
Green tea
Fish
Small fish
Natto (fermented soybean)
Miso soup
Tofu (bean curd)
Japanese confections
Cake
Peanut
Orange
Fruit
Coffee
Milk
Yogurt
Cluster in J-MICC/HERPACC
Desserts and Beverages
Meat
Seafood
Vegetable
Japanese Cuisine (Washoku)
Delta area
Relative change in area under CDF curve
Number of clusters
b
Japanese tea
Milk
Dried fish and salted fish
Salted fish roe
Miso soup
Rice
Egg
Fresh Fish
Green vegetable
Seawed
Bacon, ham, and sausage
Pork
Chick
Cheese
Beef
Liver
Fruit juice
Vegetable juice
Chinese tea
Black tea
Bread
Cafe
Noodle
Instant noodle
Cluster in JPHC
Beverage
Coffee, Bread, and Noodle
Meat and Cheese
Japanese Cuisine (Washoku)
Delta area
Relative change in area under CDF curve
Number of clusters
c
Pork
Lamb
Processed meat
Beef
Coffee
Poultry
Bread
Cheese
Oily fish
Non oily fish
Cooked vegetable
Raw vegetable
Cereal
Tea
Fresh fruit
Dried fruit
Cluster in HPP
Bread, Cheese, Poultry, and Coffee
Fish
Fruit, Cereal, and Tea
Red and Processed Meat
Vegetable
Delta area
Relative change in area under CDF curve
Number of clusters

Extended Data Figure 4. Consensus clustering of dietary questionnaires in independent replication cohorts.

Consensus clustering of food frequency questionnaires in J-MICC/HERPACC (a), JPHC (b), and HPP (c). Delta area plots show the area under the cumulative distribution function (CDF) curve when changing the number of clusters from two to ten.

"We further evaluated replication in the African and American populations in All of Us (N_{max} = 70,558 and 66,556, respectively) and the Israeli population in the Human Phenotype Project (HPP; N_{max} = 8,645). In the African population, three G×E interactions were significantly replicated, two of which overlapped with the shared G×E interactions between UKB1 and BBJ1 (**Ext. Data Fig. 5a and Supplementary Table 13**; $P_{G\times E} < 0.05/77/3 = 2.3\times10^{-4}$, corrected for the number of trait-locus-biobank triplets outside of the *ALDH2* locus and the number of populations). In the American population, two G×E interactions for pulse pressure (PP) were significantly replicated, both originally from UKB1 and primarily driven by G×Age. G×E interactions were not significantly replicated in the Israeli population, possibly due to its relatively small sample size. Nevertheless, we observed the top signal for one of the shared G×E interactions between UKB1 and BBJ1 (*UGT1A* gene family for T.Bil, commonly driven by G×Current-smoking and G×Age)." (**Results**, P13 line 296–307).



Extended Data Figure 5. Replication within and across populations.

a–b, Replication statuses of G×E loci other than the *ALDH2* locus (**a**) and the *ALDH2* locus (**b**). Only interactions with at least one nominally significant replication ($P_{G\times E} < 0.05$) are shown. EUR, the European population; EAS, the East Asian population; AFR, the African population; AMR, the American population. **c–f**, eGFR distributions across rs77924615 genotypes (the *UMOD* locus) and age in UKB1 (**c**), UKB2 (**d**), BBJ1 (**e**), and BBJ2 (**f**), as a representative example of cross-population shared G×E interaction. Dots, 5,000 randomly sampled individuals per genotype; lines, regression lines using all participants. **g and h**, minor allele frequencies (MAF) of the lead variants (**g**) and environmental distributions (**h**) in UKB1 and BBJ1. In **g**, cross-population shared loci are marked with crosses and labeled by nearest genes. Solid gray lines represent x=0, y=0, and y=x. **i**, number of G×E loci detected with $P_{G\times E}$ less than 5.0×10^{-8} in randomly subsampled participants in UKB1 and BBJ1. Subsample sizes: 10, 50, 100, 150, and 200 thousand in UKB1; 10, 50, and 80 thousand in BBJ1. Full-cohort results also shown (N_{mean} = 253,773 in UKB1 and 133,117 in BBJ1).

We described the overview of the replication cohorts in **Supplementary Methods**.

"**Japan Multi-Institutional Collaborative Cohort (J-MICC) Study and Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC)**

J-MICC is a multi-institutional collaboration of prospective and population-based cohorts comprising approximately 100,000 participants in Japan, launched in 2005 with the primary purpose of detecting G×E interactions of lifestyle-related diseases. Individuals aged 35 to 69 years were enrolled from 2005 to 2010 from respondents to study announcements in specified areas, inhabitants attending health checkup examinations by local governments, visitors at health checkup centers, and patients at a cancer hospital. The second assessment was conducted five years after their enrollment. HERPACC is a hospital-based cohort including participants aged 18 or older recruited from 2001 to 2013 from first-visit outpatients at Aichi Cancer Center Hospital. Aichi Cancer Center Hospital is one of the constitutive cohorts of J-MICC, and the same questionnaire was used for both J-MICC and HERPACC. After removing duplicated participants between J-MICC and HERPACC, we analyzed the participants of J-MICC and HERPACC together, which we refer to as J-MICC/HERPACC. For HERPACC, we analyzed the participants with no cancer or history of neoplasm to match the cohort to the population-based characteristics of J-MICC." (**Supplementary Methods**).

"**Japan Public Health Center-based Prospective (JPHC) Study**

The JPHC study is a population-based cohort comprising participants enrolled from residents aged 40-69 years within 11 public health center (PHC) areas nationwide between 1990 and 1994. Before conducting genetic research in the JPHC study, we obtained approval from the institutional review board of the National Cancer Centre (Approval No.: 2011-044), Tokyo, Japan, and provided eligible participants with the option to decline participation. For the present study, 13,024 participants were randomly selected as a subcohort from the JPHC study participants in 9 PHC areas, who had received a health check-up and answered a questionnaire during the baseline survey." (**Supplementary Methods**).

"**All of Us**

All of Us is a nation-wide cohort in the United States, aiming to recruit persons in demographic categories that have been and continue to be underrepresented in biomedical research.

National enrolment began in 2018, and adults 18 years and older who have the capacity to consent and reside in the United States or the United States territory at present are eligible. We analyzed the version 8 dataset of the All of Us." (**Supplementary Methods**).

"**The Human Phenotype Project (HPP)**

HPP is a prospective cohort in Israel, and the participants between the ages of 40 and 70 years were recruited from 2018 to 2021 from the self-assignment of volunteers registered to the trial website. The cohort is largely composed of Ashkenazi Jewish people. Individuals without recent antibiotic use and gastrointestinal morbidity were eligible if they had not yet been diagnosed with the clinical outcomes of interest." (**Supplementary Methods**).

"**Quality control of phenotypes and environments in the replication cohorts**

Clinical traits

We employed the same definition and quality control for clinical traits as those used in UKB1 and BBJ1. In All of Us, biomarker phenotypes were obtained from the electronic medical records, and we used the most recent records for each biomarker. When both medical records from the emergency department and those from other departments were available, we prioritized the latter. We observed frequent mislabeling of the measurement units in this cohort. For example, albumin measurements with the units of mg/mL and mg/L were of the same order of magnitude as those reported in g/dL. Therefore, we corrected the units by examining the median values for individual unit labels. The biomarker phenotypes for the other cohorts were obtained from the baseline assessment data for the main analyses, while we used the second assessment data of J-MICC for the temporal order analyses. We applied the same quality control criteria as used for BBJ and UKB. For disease statuses, we defined participants as cases if they had relevant SNOMED CT codes or self-reported disease histories in All of Us. We used self-reported disease histories for the other cohorts. For JPHC and J-MICC/HERPACC, self-reported histories were recorded for diabetes mellitus instead of T2D, and we used diabetes mellitus as a proxy of T2D for replication purposes.

Environmental factors

The environmental factors from questionnaires, along with the conversion of categorical responses into a continuous scale, were summarized in **Supplementary Table 4**. As only one environmental factor was related to physical activity in JPHC, clustering it with dietary environmental factors would lead to a cluster consisting of both physical activity and dietary factors, opposite to our observation that the "physical activities" cluster was consistently detected as a separate cluster from other dietary clusters in UKB1 and BBJ1. Therefore, we pre-defined the "physical activities" cluster from the environmental factors related to physical activities in the replication cohorts and performed clustering only using dietary factors. In JPHC, two sets of questionnaires with slight differences in questionnaire items were used, and we used the items available in both sets. Questionnaires related to physical activities and dietary consumption were unavailable in All of Us, and we did not perform clustering in this cohort. In All of Us, smoking status was separately asked for cigar smoking, electronic

smoking, hookah smoking, and smokeless tobacco smoking. We defined ever- and current-smoking statuses based on those of all types of tobacco." (**Supplementary Methods**).

We added the results of the subsampling analysis to the **Results** section as follows.

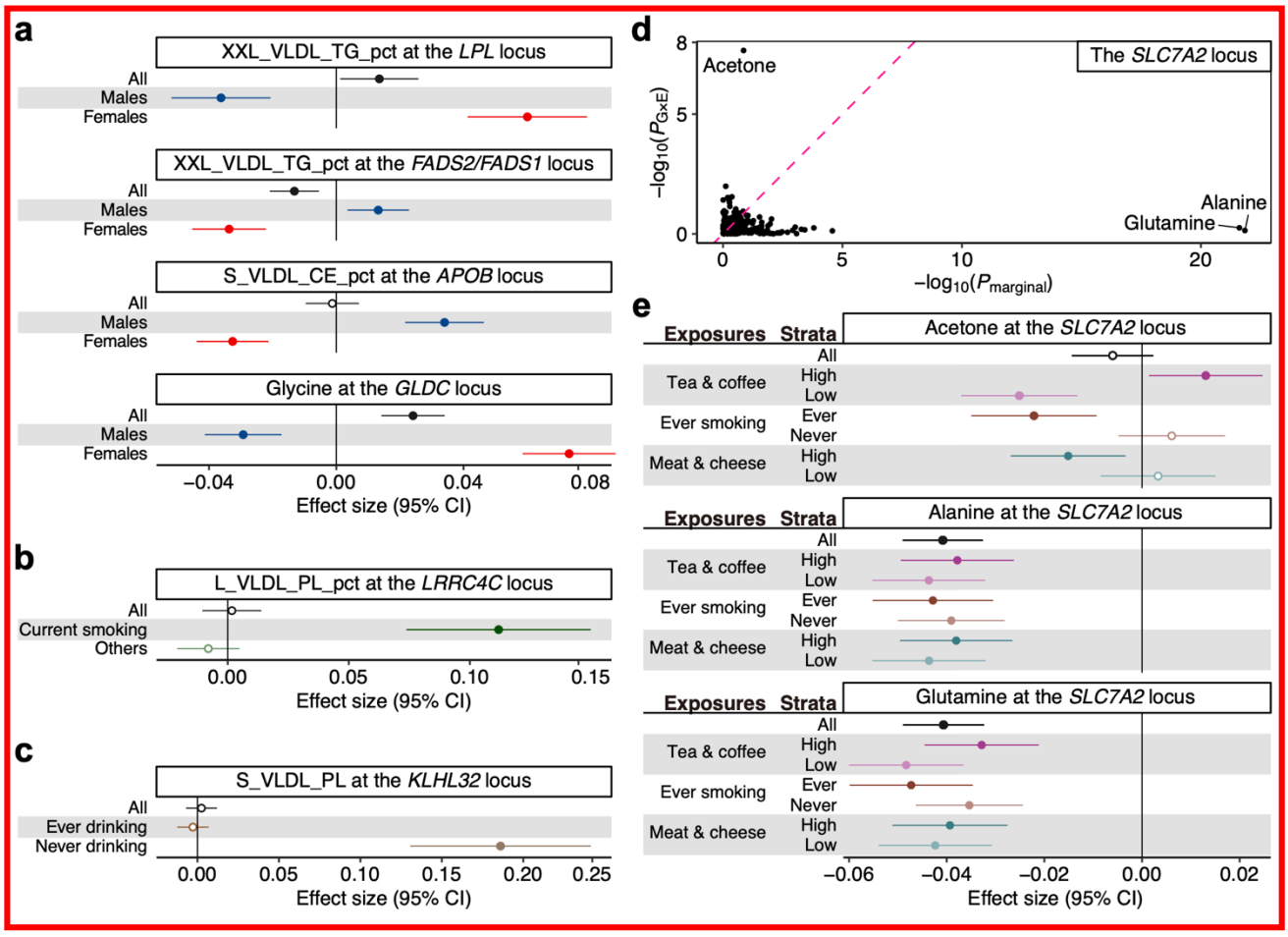
"Random subsampling analyses revealed that detecting G×E interactions required biobank-scale sample sizes, and detected loci were not saturated (**Ext. Data Fig. 5i**), indicating the need for biobank collaboration across diverse populations to detect additional G×E interactions." (**Results**, P13 line 292–295).

We also added the genome-wide G×E interaction tests for the metabolome in the **Results** section.

"Motivated by these G×E metabolite QTLs, we expanded the metabolome-wide G×E analysis to genome-wide variants, identifying 30 and 11 genome-wide significant loci in UKB1 and BBJ1, respectively, yielding 736 and 228 metabolite–locus pairs. Among these, 16 loci in UKB1 and 4 in BBJ1 passed the Bonferroni threshold ($P_{G\times E} < 7.7\times10^{-11}$, correcting for 325 metabolites across two cohorts). Notably, G×Sex interactions at the *ALDH1A2* and *ZNF259* loci were observed in both cohorts (**Supplementary Table 31**). Several loci exhibited sex-dimorphic G×Sex interactions with opposite effect directions in males and females. These included the *LPL*, *FADS2/FADS1*, *APOB*, and *GLDC* loci (**Ext. Data Fig. 10**; $P_{G\times E} = 8.7\times10^{-24}$, 7.7×10^{-27}, 6.6×10^{-17}, and 4.9×10^{-28}, respectively), again underscoring the sex-specific genetic architecture of metabolome regulation. Additional notable G×E signals, though below Bonferroni significance, included a current-smoker-specific association at the *LRRC4C* locus for phospholipids to total lipids ratio in large VLDL (L_VLDL_PL_pct; $P_{G\times E} = 1.1\times10^{-8}$), and a never-drinker-specific association at the *KLHL32* locus for phospholipids in small VLDL (S_VLDL_PL; $P_{G\times E} = 4.8\times10^{-9}$). At the *SLC7A2* locus, we observed marginal effects for alanine and glutamine and a G×E-only effect for acetone ($P_{G\times E} = 2.2\times10^{-8}$), driven by tea and coffee consumption, ever smoking, and meat and cheese consumption (**Ext. Data Fig. 10**). Further validation and functional studies are warranted to elucidate their underlying mechanisms." (**Results**, P25–26 line 615–632).

"We further conducted omics-level G×E QTL analyses to pinpoint the metabolites and proteins that mediate clinical G×E interactions. Multiple sex-dimorphic and environment-specific genetic effects on the metabolome suggest that omics data may reveal G×E interactions with large effect sizes." (**Discussion**, P27–28 line 660–665).

"For the genome-wide G×E interaction tests for metabolome, we targeted (i) genotyped variants in each cohort, (ii) HapMap3 variants, and (iii) the variants targeted by the meta-analysis across BBJ and UKB, to reduce the computational cost. After defining the metabolome G×E loci based on their distance as above, we merged the metabolome G×E loci if they overlapped with the same locus defined for the clinical traits." (**Methods**, P44–45 line 1124–1129)



Extended Data Figure 10. Genome-wide G×E interactions for metabolites.

a, sexual dimorphic effects at four loci in UKB1. XXL_VLDL_TG_pct, the triglycerides to total lipids ratio in chylomicrons and extremely large VLDL; S_VLDL_CE_pct, the cholesteryl esters to total lipids ratio in small VLDL. **b and c**, environment-specific effects in UKB1 for current-smoking at the *LRRC4C* locus (**b**) and never-drinking at the *KLHL32* locus (**c**). L_VLDL_PL_pct, the phospholipids to total lipids ratio in large VLDL; S_VLDL_PL, phospholipids in small VLDL. **d**, distribution of $P_{marginal}$ and $P_{G\times E}$ at the *SLC7A2* locus in UKB1. **e**, environment-stratified effect sizes at the *SLC7A2* locus in UKB1. For the "tea and coffee" and "meat and cheese" environment clusters, participants were stratified into two equal-sized groups. Note that coffee consumption was multiplied by -1 to account for its negative correlation with tea consumption in UKB1; therefore, higher cluster values indicate more frequent tea consumption, whereas lower values indicate more frequent coffee consumption.

2. The approach for identifying significant GxE is confusing and unclear. Lines 217 – 225, and lines 911- 940: it appears that an E was considered significant it if it improved the overall significant model, but not based on whether the interaction itself was significant. Please clarify whether Es were also examined independently before being considered significant. There is

a big difference in interpretation if an individual interaction term is significant versus the overall model fit is improved due to the inclusion of an interaction term.

Response:

Thank you for your comment. We agree with you that a clearer representation of significant G×E interactions would be beneficial in further enhancing readability. In our manuscript, we performed multiple G×E interaction tests: (i) testing each single G×E term and (ii) testing multiple G×E terms related to multiple environments jointly. **We accounted for multiple testing of environment sets by combining the per-environment *P*-values into a single variant-level *P*-value using the Cauchy combination method.** The Cauchy combination method has been reported as accurate under arbitrary dependency structures (Liu Y and Xie J, ***J Am Stat Assoc*** (2021)) and has been widely adopted in genetic studies (Li X *et al.* ***Nat Genet*** (2020), for example).

We defined a G×E interaction as significant at the variant level if the combined *P*-value was below the genome-wide threshold of 5.0×10^{-8}. We also reported the study-wide Bonferroni threshold, as described below. We did not require any individual environment's interaction term to be significant on its own. Nevertheless, we note that **at least one per-environment *P*-value was smaller than the Cauchy-combined *P*-value in principle**, as the Cauchy combination method is not a meta-analysis but rather returns a *P*-value within the range of per-environment *P*-values. We have reported both Cauchy-combined and per-environment *P*-values in **Supplementary Table 6** for full transparency in both the initial and revised manuscripts. In addition, we will release genome-wide summary statistics for all possible combinations of phenotypes and environments, further ensuring full transparency.

Our approach allowed us to capture signals arising from multiple environments, in addition to those relying on a single strong signal, which was particularly relevant for loci such as *FTO* and *ALDH2*. We note that the approach to jointly test multiple G×E terms has been established in Moore R *et al.* ***Nat Genet*** (2019). They reported that the *FTO* locus surpassed the genome-wide significance threshold only when testing multiple G×E terms, and testing a single G×E term individually would have missed this locus.

After evaluating the significance of G×E interactions at the variant level, we employed an iterative approach to determine the environmental factors that contributed to the G×E interactions. Briefly, we added the G×E interaction terms one by one to the regression model in the order of their importance if the model fit was improved with a P-value less than 0.05 (likelihood ratio test). In principle, this approach selected more parsimonious sets of environments than the Akaike information criterion and can determine the top environment contributing to individual G×E interactions. Multiple environments were chosen if they remained informative for explaining the G×E interactions when conditioning on the environments already selected in previous iterations.

Our approach to report locus-level *P*-values using the Cauchy combination method was closely related to the multiple testing correction that you mentioned in #3. **We did not need to correct for the number of environments as we combined per-environment *P*-values into a single per-variant *P*-value.** We reported both the genome-wide significant threshold (5.0×10^{-8}) and the cohort-wide Bonferroni threshold (5.0×10^{-8} / 47 traits) in the initial manuscript. Following your thoughtful suggestion in Comment #3, we have now accounted for the number of cohorts for the Bonferroni-corrected *P*-value threshold (5.0×10^{-8} / 47 traits / 2 cohorts; the study-wide Bonferroni threshold). This threshold was conservative, as we found that some G×E interactions were shared between populations. We appreciate your suggestion, which helped us improve the clarity and transparency of our significance criteria.

Modification:

According to your kind suggestion, we revised the **Methods** section to further enhance readability.

"**G×E interaction testing**

All statistical tests were two-sided unless otherwise noted. As phenotypes, we targeted quantitative traits (clinical biomarkers, metabolites, and protein expression measurements), dichotomous traits (disease statuses), and time-to-event traits (disease onset for incidental cases and overall survival). For quantitative and dichotomous traits, we tested G×E interactions using GEM, a fast and scalable implementation of linear and logistic regressions with G×E interaction terms[23] (**Ext. Data Fig. 1**). We estimated the model-robust "sandwich" standard errors to suppress the inflation of statistics[23]. As the case–control imbalance can cause the inflation of test statistics in the logistic regression, we re-estimated the effect sizes and *P*-values using Firth logistic regression for the variants with *P*-values less than 5.0×10^{-8} for dichotomous traits. For the time-to-event traits, we employed the Cox proportional hazard model implemented in the R package "survival". Individuals who had already developed the target disease at baseline or developed it within six months from baseline were excluded from the incidental case analyses.

We tested G×E interactions for different sets of environmental factors, as shown in **Ext. Data Fig. 1**, and then aggregated *P*-values across environmental sets on the Cauchy distribution[31] to obtain per-variant *P*-values. When *P*-values were 0, which meant *P*-values were less than 1.0×10^{-300}, we estimated the precise *P*-values using the multiple-precision floating-point arithmetic with the maximal precision of 10,000 bits, implemented in the Rmpfr R package. We defined a G×E interaction as significant at the variant level if the aggregated *P*-value was below the genome-wide threshold of 5.0×10^{-8}. We also reported the study-wide Bonferroni threshold for full transparency. We did not require that any individual environment's interaction term be significant on its own, considering that G×E interactions can be driven by multiple environments. Nevertheless, we note that at least one raw *P*-value was smaller than the Cauchy-combined *P*-value in principle, as the Cauchy combination method is not a meta-analysis but rather returns a *P*-value within the range of raw *P*-values.

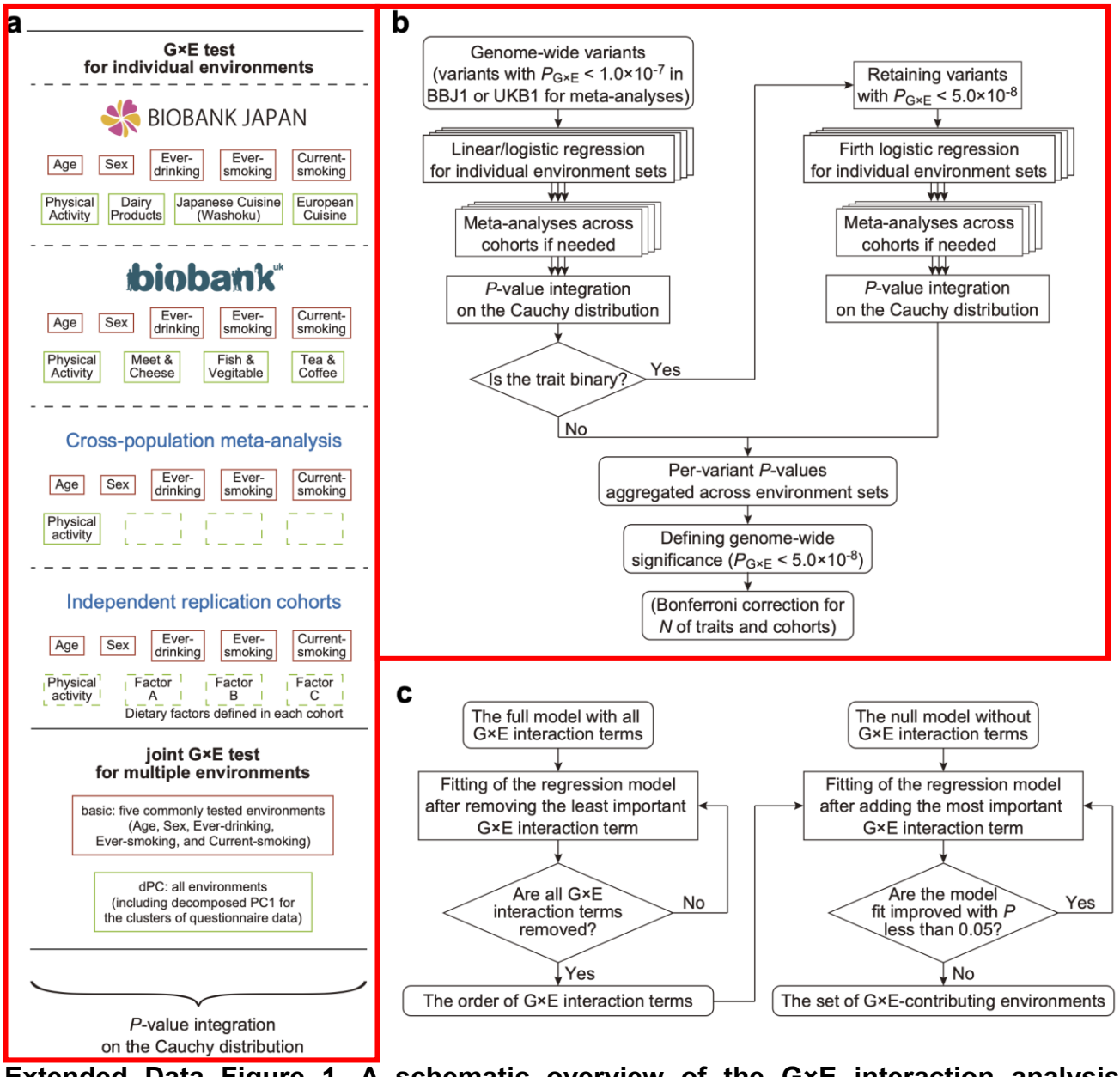
After evaluating the significance of G×E interactions at the variant level, we determined the order of importance of G×E interaction terms by backward elimination from the regression model with all G×E interaction terms. Specifically, we repeatedly removed the G×E interaction term with the least improvement in the likelihood of the linear regression model or the largest Wald *P*-value of the Firth logistic regression model. After removing all G×E interaction terms, we brought back the G×E interaction terms one by one in the order of their importance if the model fit was improved with a *P*-value less than 0.05 (likelihood ratio test). We considered that the environmental factors contributed to G×E interactions if the corresponding interaction terms were included in the final model. In principle, this approach selects more parsimonious sets of environments than the Akaike information criterion and can determine the top environment contributing to individual G×E interactions.

We used the first 20 genetic principal components (PC), genotype array, age^2, age × sex, and age^2 × sex as covariates for UKB. For the other cohorts, we used the first 10 PCs, age^2, age × sex, and age^2 × sex as covariates. Age and sex, as well as other environmental factors,

were included as the interaction terms with genotypes, and the variables used for interaction terms were automatically included as covariates. We also included the status of the target diseases in BBJ as covariates for quantitative traits and overall survival. We included the status of statin medication as an additional covariate for metabolome measurements. For sex-specific diseases in PheWAS, we excluded sex from the environmental factors and age × sex and age^2 × sex from covariates.

Following that previous cross-population GWAS used the distance-based locus definition[22], we defined that the genome-wide significant variants were in the same locus if their distance was less than 500 kbp. For the genome-wide G×E interaction tests for metabolome, we targeted (i) genotyped variants in each cohort, (ii) HapMap3 variants, and (iii) the variants targeted by the meta-analysis across BBJ and UKB, to reduce the computational cost. After defining the metabolome G×E loci based on their distance as above, we merged the metabolome G×E loci if they overlapped with the same locus defined for the clinical traits." (**Methods**, P43–45 line 1076–1129).

We also revised **Ext. Data. Fig. 1** (**Supplementary Figure 1** in the initial manuscript) to represent our approach more clearly.



Extended Data Figure 1. A schematic overview of the G×E interaction analysis workflow.

a, The environment sets used to test G×E interactions. *P*-values were calculated for individual environmental sets and integrated on the Cauchy distribution to obtain final *P*-values. **b**, The workflow of G×E interaction tests. **c**, the workflow of defining the environments which contributed to G×E interactions and their orders.

We note that we briefly described our approach in the first section of **Results** in both the initial and revised manuscripts.

"For each trait and in each cohort, we tested G×E interactions for nine environmental factors both individually and jointly to boost power and aggregated *P*-values into one value per variant on the Cauchy distribution (**Ext. Data Fig. 1**)." (**Results**, P8 line 165–167).

We also reported a study-wide Bonferroni threshold along with the genome-wide significant threshold, instead of the cohort-wide one in the initial manuscript, by additionally accounting for the number of discovery cohorts. We also clarified in the **Results** section that multiple testing correction for the number of environments was unnecessary, as we reported the aggregated *P*-values across environments.

"Applying a stringent study-wide Bonferroni correction across 47 traits and two cohorts, 30 interactions in 23 loci remained significant ($P_{G\times E} < 5.3\times10^{-10}$). As *P*-values were aggregated across environments, no further correction for the number of environments was required." (**Results**, P9 line 182–185).

"In BBJ1, we observed 36 significant G×E interactions in 15 loci (26 in 8 loci after Bonferroni correction, $P_{G\times E} < 5.3\times10^{-10}$) " (**Results**, P9 line 200–201).

3. The study-wide significance threshold does not account for the number of environmental interactors in addition to the number of traits. The appropriate correction is something closer to 5x10-8/(47 phenotypes * 12 environments) = 8.8x10-11. You could also argue 2 discovery cohorts and an additional joint environment model could make this even more stringent. Although hard to tell from ST5, this takes significant loci down to something closer to 18. These strict multiple testing burdens are why few GxE have been discovered in the first place. Furthermore, it is known that a genome-wide gene-environment interaction test has better power than a vqtl prioritization approach when the E is known. (Pare et al, Plos Genet 2010), however, iterating through multiple environments immediately and harshly diminish power due to multiple testing. I would argue only findings that surpass this more stringent threshold be presented, or at the very least, some clearer communication about this rationale.

Response:

Thank you for your comment. As described in our response to Comment #2, **we accounted for multiple testing of environments by combining per-environment *P*-values into a single variant-level *P*-value using the established Cauchy combining method**. We reported both the conventional genome-wide significant threshold and the cohort-wide Bonferroni threshold in the initial manuscript. In the revised manuscript, **we changed the latter to the study-wide Bonferroni threshold by additionally accounting for the number of cohorts ($5.0 \times 10^{-8} / 47 / 2$)**. At this more stringent and conservative threshold, 30 (for UKB1) and 26 (for BBJ1) G×E interactions were significant.

The use of the genome-wide significant threshold was consistent with previous GWAS studies targeting multiple phenotypes (Zhou W *et al.* ***Cell Genom*** (2022) and Sakaue S and Kanai M *et al.* ***Nat Genet*** (2021)). We also note that the genome-wide significant threshold was also employed without correcting for the number of phenotypes in Bernabeu *et al.* ***Nat Genet*** (2021), which you kindly introduced in Comment #6. Pleiotropy has also been evaluated using this threshold (Watanabe K *et al.* ***Nat Genet*** (2019)), successfully yielding biological insights.

We agree with you that multiple testing burdens were one of the reasons why few G×E have been discovered previously. We would like to clarify that researchers have discussed

the multiple testing burdens to correct for the number of environments (Dick DM *et al.* ***Perspect Psychol Sci*** (2015), Joseph PG *et al.* ***Can J Cardiol*** (2013)). Multiple testing correction for the number of phenotypes and cohorts is not unique to G×E studies, but generally relevant to phenome-wide and cross-population studies. In line with previous cross-population GWAS studies targeting multiple phenotypes (Zhou W *et al.* ***Cell Genom*** (2022) and Sakaue S and Kanai M *et al.* ***Nat Genet*** (2021)), we adopted the genome-wide threshold (5.0×10^{-8}) for main discovery, while also reporting the stricter Bonferroni threshold for full transparency.

We also thank you for introducing Pare *et al.*, which was consistent with our observation that vQTL prioritization approach could fail to capture G×E interactions at both variant and genome-wide levels. This underscores the importance of direct genome-wide G×E testing, as conducted in our study. We also agree with you that clearer communication about the rationale for the significant threshold would be helpful for readers. We have revised the manuscript accordingly, as described in #2.

Modification:

Please see #2 for the modification.

4. Similar to #3, I find the replication P-value correction insufficient. The authors define significant by the number of SNPs tested in each cohort, but in reality, they are testing 98 total SNPs for replication (no matter which cohort).

Response:

Thank you for your thoughtful comment. Following your suggestion, we used a replication threshold of 0.05 / 98 based on the 98 trait–locus pairs detected in BBJ1 and UKB1.

Modification:

We updated the number of replicated G×E interactions using the revised threshold.

"Out of the 62 G×E interactions in UKB1, 23 showed nominal replication in UKB2 ($P_{G\times E}$ < 0.05), and 7 were significantly replicated after Bonferroni correction ($P_{G\times E} < 5.1\times10^{-4}$; corrected for 98 G×E interactions across UKB1 and BBJ1) (**Figure 1c and Supplementary Table 6**). In BBJ, 28 out of 36 G×E interactions were nominally replicated, and 19 were significantly in BBJ2 (**Figure 1e and Supplementary Table 6**)" (**Results**, P11 line 230–235).

5. It is most appropriate to only discuss loci that surpass both a strict study-wide significant threshold and replicate (at least in the same biobank). Therefore, I recommend at least mentioning these limitations (replication and p-value) throughout the manuscript when loci that did not replicate are highlighted. Furthermore, more discussion around the lack of replication is warranted.

Response:

Thank you for your comment. As noted in our responses to Comments #2 and #3, we followed the convention of previous GWAS studies targeting multiple phenotypes to adopt the genome-wide threshold and also reported the study-wide Bonferroni threshold for full transparency. As we wrote in response to Comment #2, pleiotropy has been evaluated using the genome-wide significance threshold (Watanabe K *et al.* ***Nat Genet*** (2019)), successfully yielding its biological insights. Indeed, the genome-wide significant loci provided an overview of G×E interactions in our manuscript, including pleiotropy and cross-population consistency.

In addition, **we focused on the G×E interactions at the *ABCG2*, *ALDH2*, and *PITX2* loci in the manuscript, all of which surpassed the study-wide Bonferroni threshold**. The *ABCG2* locus was nominally replicated, and the *ALDH2* locus was significantly replicated. Curiously, the *PITX2* locus lacked replication, which we discussed in the section on the G×Natto interaction. **We reported their *P*-values in the discovery cohort and their replication statuses in the initial manuscript**. In the revised manuscript, we additionally conducted metabolome-wide G×E interaction tests, as described in our response to Comment #1. In this analysis, we highlighted G×Sex interactions that surpassed the study-wide Bonferroni threshold. We also described other genome-wide significant G×E interactions for metabolome, while ensuring to note that these G×E interactions did not surpass the study-wide Bonferroni threshold in the manuscript.

As described in response to Comment #1, we have extensively evaluated replication using UKB2 and BBJ2, as well as independent cohorts (N = 436,272 from four cohorts). Our analysis represents the largest effort to date in providing a list of robust G×E interactions. Replication in diverse populations also provided a list of cross-population shared G×E interactions.

The replication rates of G×E interactions were 11.3% (7/62) in UKB2 and 52.8% (19/36) in BBJ2, at the Bonferroni-corrected significance threshold. The rate was higher in BBJ2, possibly due to its larger number of participants than UKB2. To empirically evaluate these rates in comparison to those of GWAS, we newly performed GWAS for the traits with genome-wide significant G×E interactions in UKB1 and BBJ1. When using all GWAS loci that surpassed the Bonferroni-corrected threshold, the replication rates were 10.0% (522/5,238) in UKB2 and 34.0% (613/1,802) in BBJ2, which were lower than those of G×E interactions. In this analysis, the Bonferroni-corrected thresholds were different between G×E and GWAS. Therefore, we next filtered the GWAS trait–locus pairs to match the Bonferroni-corrected thresholds. Specifically, for each G×E locus, we selected the marginal effect locus for the same trait whose *P*-value was closest to that of the corresponding G×E interaction. Using this matched set, the replication rates of GWAS were only slightly higher than those of G×E (14.5% (9 / 62) in UKB2 and 66.7% (24 / 36) in BBJ2). In summary, **these additional results demonstrate that the replication rates of G×E interactions reported in our manuscript were largely comparable to those of GWAS**. Considering that differences in environmental distributions across cohorts can reduce the replicability of G×E interactions, and that previous G×E studies have been criticized by their low replication rates, these findings suggest the high reliability of the G×E interactions identified in this manuscript.

Modification:

We added the comparison of replication rates between G×E interactions and marginal (GWAS) effects in **Results and Supplementary Note 4**.

"The replication rates were largely comparable to the GWAS of the same trait sets, supporting the reliability of the G×E interactions that we reported (**Supplementary Note 4**)." (**Results**, P11 line 243–245).

"**Supplementary Note 4. Comparable replication rates of G×E interactions compared to those for GWAS**

To empirically compare the replication rates of G×E interactions to those of GWAS, we performed GWAS for the traits with genome-wide significant G×E interactions in UKB1 and BBJ1. We detected 5,238 and 1,802 trait–locus pairs with genome-wide significant marginal effects in UKB1 and BBJ1, respectively (7,040 in total). In UKB2, 10.0% (522 / 5,238) were significantly replicated at the Bonferroni-corrected threshold of 7.1×10^{-6} (= 0.05 / 7,040), and 34.0% (613 / 1,802) were significantly replicated in BBJ2, reflecting its larger sample size than UKB2. The replication rates of G×E interactions were higher than these rates: 11.3% (7 / 62) in UKB2 and 52.8% (19 / 36) in BBJ2 at the Bonferroni-corrected threshold of 5.1×10^{-4} (= 0.05 / 98).

In the above analysis, the Bonferroni-corrected thresholds were different between G×E and GWAS. Therefore, we next filtered the GWAS trait–locus pairs to match the Bonferroni-corrected thresholds. Specifically, for each G×E locus, we selected the marginal effect locus for the same trait whose *P*-value was closest to that of the corresponding G×E interaction. Using this matched set, the replication rates of GWAS were 14.5% (9 / 62) in UKB2 and 66.7% (24 / 36) in BBJ2, only slightly higher than those of G×E interactions.

In summary, we observed that the replication rates of G×E interactions were largely comparable to those of GWAS. Considering that differences in environmental distributions across cohorts can reduce the replicability of G×E interactions, and that traditional G×E studies have suffered from low replication rates, these findings suggest the high reliability of the G×E interactions identified in this manuscript." (**Supplementary Note 4**).

6. Unfortunately, the GWAS catalog is not a reliable resource for classifying loci or trait-loci pairs in a GxE study as known versus novel given so few of these findings have made it to that resource. A quick check, and I found at least 7 trait-gene pairs that have already been identified in the same exact dataset, UK Biobank (APOE-lipids, LEPR-CRP, ALPL-ALP, PNPLA3-AST, CRP-CRP, CPS1-Plt, SLC2A9-Urate/Gout; Bernebeu et al Nat Genet 2021, Westerman et al Nat Commun 2022). Many of your signals are driven by Sex and the Bernebeu et al manuscript tested UKB-wide for gene-sex interactions on >500 traits, and similarly, the Westerman et al manuscript tested UKB-wide for GxE with >2000 environmental interactors on blood biomarkers. A more thorough look at what has already been done in UKB or BBJ in the GxE space would give a more accurate answer to the question of novelty. The true novelty of this manuscript is not in the GxE discovery, but rather the deeper investigation at top loci.

Response:

Thank you for your helpful comment and for kindly pointing us to the important studies by Bernabeu *et al.* (***Nat Genet*** (2021)) and Westerman *et al*. (***Nat Commun*** (2022)). We agree

with you that the GWAS catalog is not a comprehensive resource for G×E interactions. The limited coverage of G×E studies in public databases underscores the need for systematic cataloging efforts, which we aimed to contribute to through our genome-wide analyses targeting a large number of combinations of phenotypes and environments in the European and East Asian populations. Importantly, we avoided using the term “novel” in both the initial and revised manuscript, in line with Nature group’s editorial policy.

For the GWAS catalog, we also noticed that some studies that did not investigate G×E interactions (e.g., studies focusing on gene-by-gene [G×G] interactions) were incorrectly classified as G×E studies. Nevertheless, **when we filtered the GWAS catalog by partial matching of the trait names and manually revised all studies that passed the filter, we found that the GWAS catalog included 76 studies that reported G×E interactions with *P*-values less than 5.0×10^{-8}**. These studies reported 217 G×E interactions for 203 loci in total. As a starting point, we examined the locus-level overlap between the G×E interactions we reported and those in the GWAS catalog. Among the 53 loci that we reported, we detected one overlapped locus with a matched trait and environment (G×Current-smoking in the *HYKK* locus for BMI), which we reported in the initial manuscript. We have now compiled the list of curated G×E interactions from the GWAS catalog as **Supplementary Table 7** to aid future efforts in validating G×E interactions.

We also examined the locus overlap with the two reports you kindly introduced. Bernabeu *et al* investigated G×Sex interactions with a different statistical model from our study, and Westerman *et al* investigated G×E interactions for cardiometabolic serum biomarkers after filtering the variants to the variance quantitative loci (vQTL). Due to the pre-filtering of target variants, Westerman *et al* applied a more lenient significance threshold (2.35×10^{-7}) compared to ours. In addition, these two studies reported per-environment *P*-values, whereas we defined significance based on locus-level, environment-aggregated *P*-values, which were inherently more stringent at the same threshold. We observed 7 loci (11 trait–locus pairs) and 12 loci (15 trait–locus pairs) overlapped, respectively. These loci included all overlaps you kindly highlighted and supported the validity of our findings. As both Bernabeu *et al.* and Westerman *et al.* only analyzed the UKB data, our assessment of replication in independent cohorts provides validation for these overlapping loci. The replication in independent cohorts is particularly important for G×E interactions, as the previous G×E studies have been criticized for their low replication rates.

Since Bernabeu *et al* focused exclusively on G×Sex interactions, it is worth noting that the top contributor of the overlapping G×E interactions was G×Sex for all loci (7 out of 7 loci; 9 out of 11 trait–locus pairs across UKB1 and BBJ1). In our manuscript, G×Sex was the top contributor of G×E interactions at 36 loci in UKB1 and 11 loci in BBJ1, and only a small fraction of them overlapped with Bernabeu *et al* (5/36 for UKB1 and 4/11 for BBJ1). This suggests that including a broader range of environmental factors allowed us to detect G×E interactions beyond those previously identified through the G×Sex-only approach. Of note, the other G×E interactions detected in our manuscript were primarily driven by environments other than sex.

In addition to the above overlaps, we have reported two loci previously reported in the initial manuscript: the G×Age interaction in the *UMOD* locus for eGFR and the G×E interaction in the *FTO* locus for BMI driven by multiple environments. As previous studies have focused primarily on the European population, G×E interactions have rarely been established in the East Asian population. Although not genome-wide significant, our co-authors first reported G×Drinking interactions at the *ALDH2* locus for esophageal cancer in 2001 (Matsuo K *et al.* ***Carcinogenesis*** (2001)). We confirmed this G×E interaction with

genome-wide significance ($P_{G\times E} = 1.2\times10^{-32}$) in the PheWAS section and referred to Matsuo K *et al.* in both the initial and revised manuscripts. BBJ has joined the CHARGE consortium to explore G×E interactions for serum lipids (Bentley AR *et al.* ***Nat Genet*** (2019)). This study reported three genome-wide significant G×E loci, all of which were detected only in the African population in their study and did not overlap with the loci reported by us. Their East-Asian-specific analysis did not yield a genome-wide significant locus. We note that only a subset of BBJ participants were analyzed in their study, which may explain why the loci reported by us were missed in their study.

In total, we observed overlaps with previous reports for 32% (29/92) of the G×E interactions. We believe that this level of overlap validates our methodology through consistency with prior works and demonstrates the value of our effort to expand the catalog of G×E interactions.

Modification:

We added the description of curating the GWAS catalog and the overlap with previous studies using UKB in **Results and Supplementary Notes 2 and 3**.

"The others have not been reported with $P_{G\times E}$ less than 5.0×10^{-8} based on our curation of the GWAS catalog using trait-name matching (**Supplementary Note 2 and Supplementary Table 7**). We also summarized 16 loci overlapping with recent UKB G×E reports using different protocols in **Supplementary Note 3 and Supplementary Table 8**." (**Results**, P9 line 190–194).

"**Supplementary Note 2. Curation of previously reported G×E interactions from the GWAS Catalog**

The GWAS catalog includes a substantial number of gene–environment (G×E) studies. Although it may not encompass all such studies, it serves as a valuable resource for examining previously reported G×E interactions. To assess the overlap between the G×E interactions identified in our manuscript and those previously reported, we curated and compiled a list of G×E interactions from the GWAS catalog.

The GWAS catalog provides a specific flag "GXE" to denote G×E studies, which were marked for 268 studies (795 study–trait pairs with $P < 5\times10^{-8}$) as of July 8, 2024. However, a closer inspection revealed that not all flagged entries represented true G×E studies; some were gene–gene (G×G) interaction studies, while others lacked the "GXE" flag despite focusing on G×E interactions.

To define a reliable set of previously reported G×E interactions, we filtered the studies reporting genome-wide significant variants ($P < 5\times10^{-8}$), using partial matches in trait names followed by manual review of all retained studies, according to the criteria below:

1. Putative G×E interactions:
    (i) Included the term "interaction" or "Interaction"
    (ii) Did not match the regular expression "adjusted for .+ interaction", which meant the phenotype of standard GWAS was adjusted for some interaction term
    (iii) Did not contain the phrases indicating G×G interactions or joint tests of G×E interactions and a marginal effect: "SNP x SNP interaction", "4df test", "2df test",

"(2df)", "APOE e4 interaction", "x APOL1 genotype interaction", "GSK3-beta interaction protein levels", "SNP x mitochondrial", "joint analysis main effects and ", "(main effect)", "Stromal interaction molecule 1 level", "(3df)"

2. Putative joint tests of G×E interactions and a marginal effect:
   (i) Included the term “interaction” or “Interaction”
   (ii) Did not match the regular expression "adjusted for [^¥(]+ interaction"
   (iii) Contained at least one of the following phrases: "4df test", "2df test", "(2df)", "(3df)", "joint analysis main effects and", "joint analysis for main effect and"
   (iv) Did not include "rs671", which indicated G×G interactions

Using these criteria, we initially identified 103 studies (178 study–trait pairs) as putative G×E interaction studies, and 32 studies (89 study–trait pairs) as putative joint test studies. We then manually reviewed each study, and the following studies were excluded:

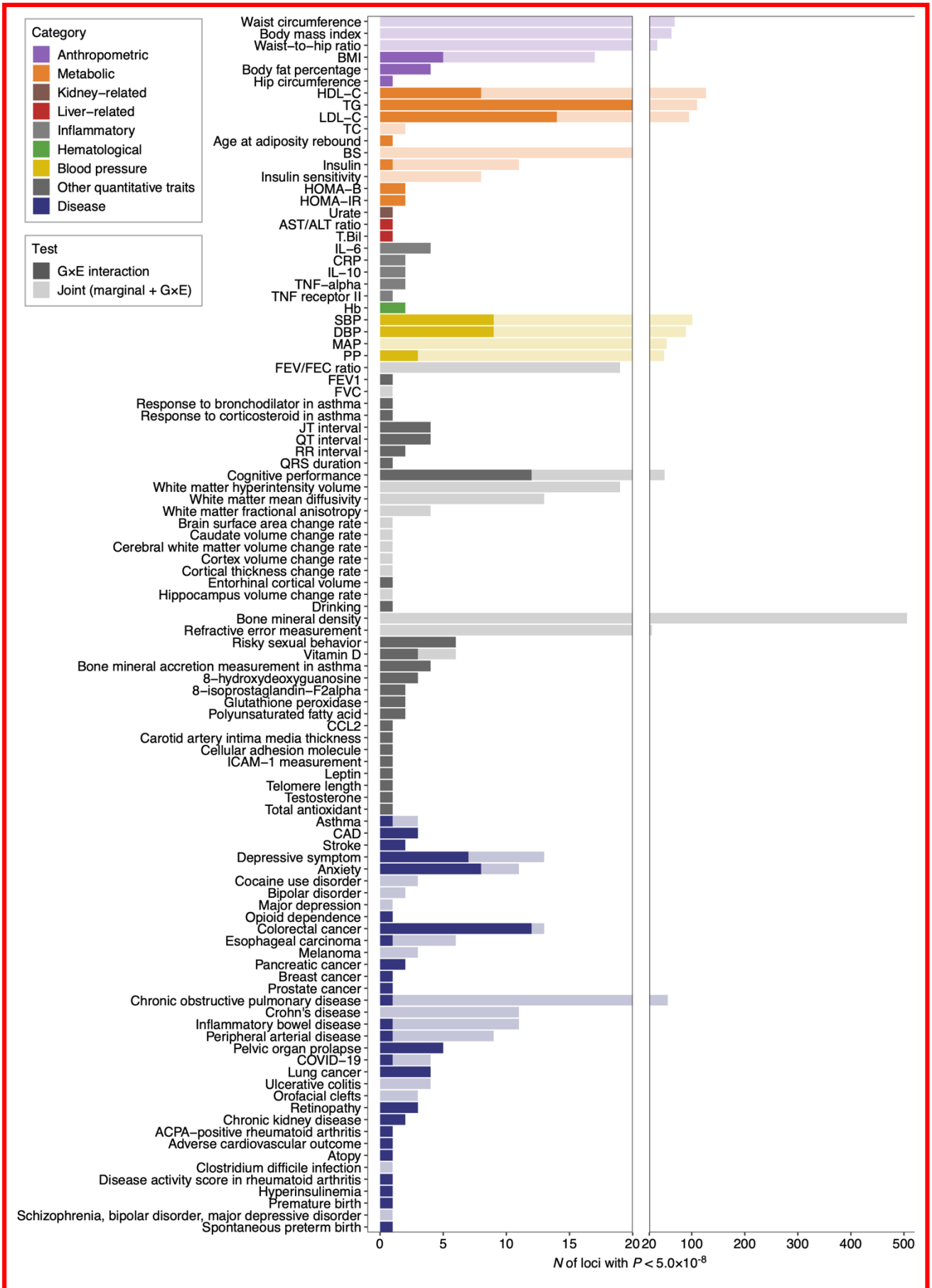
(i) Studies that did not evaluate multiplicative G×E interactions

(ii) Studies suspected of inflated results, specifically, those reporting genome-wide significant G×E loci from cohorts with only several hundred participants, or studies with no more than 2,000–3,000 participants reporting over 10 genome-wide significant G×E loci, >80% of which included only one significant variant per locus.

We also manually curated phenotype names from the GWAS catalog, as the phenotype names for G×E interactions were not standardized, often including phenotype and environment names and degrees of freedom for statistical tests.

After manual review, some entries initially categorized as G×E interaction tests were reclassified as joint tests. Ultimately, we identified 76 studies reporting G×E interactions for 203 loci (217 trait–locus pairs) and 46 studies reporting joint tests for 998 loci (1571 trait–locus pairs) **(Supplementary Figure 1)**. We found that pleiotropic G×E effects across multiple traits were rarely reported in this curated list, possibly due to the lack of systematic evaluation across phenotypes and environments.

In summary, we systematically curated and compiled G×E interactions from the GWAS catalog, providing a rigorously reviewed resource to support future validation and discovery efforts in G×E research. The full curated list is available as **Supplementary Table 7**.



Supplementary Figure 1. G×E interactions curated from the GWAS catalog.

Curated G×E interactions and joint tests of G×E interactions and a marginal effect with *P*-values less than 5.0×10^{-8}." (**Supplementary Note 2**).

"**Supplementary Note 3. Overlaps of G×E interactions with recent UKB studies**

We examined the overlap of the G×E interactions detected in this manuscript with two recent studies using UKB: Bernabeu E *et al.* and Westerman KE *et al*. The former focused on G×Sex interactions, with a different statistical model from our study. The latter investigated G×E interactions for cardiometabolic serum biomarkers after filtering the variants to the variance quantitative trait loci (vQTL). We note that these two studies reported per-environment *P*-values, whereas we defined significance based on locus-level, environment-aggregated *P*-values, which were inherently more stringent at the same threshold. In addition, Westerman *et al*. applied a more lenient significance threshold (2.35×10^{-7}) compared to ours, as they adopted a two-step approach using the vQTL filtering.

We regarded the G×E loci as overlapping when the lead variants of the previous studies were located within 500 kb of the genome-wide significant variants detected in this manuscript. To account for differences in the trait definition, we considered trait–locus pairs as overlapping if the traits belonged to the same category (e.g., uric acid and gout). Moreover, overlaps were counted regardless of differences in the environmental factors tested.

For the G×E interactions we detected in UKB1, we found 4 overlapping loci (6 trait–locus pairs) for Bernabeu E *et al.* and 12 overlapping loci (14 trait–locus pairs) for Westerman KE *et al*. (**Supplementary Table 8**). For those detected in BBJ1, five loci (five trait–locus pairs) were overlapped with Bernabeu E *et al.*, and one locus (one trait–locus pair) was overlapped with Westerman KE *et al*. These loci included the *PITX2* locus for arrhythmia, the *CYP19A1* locus for height, and the *SLC22A11* locus for UA, all of which were overlapped with Bernabeu E *et al.* These loci would be additional candidates of cross-population shared G×E interactions, although further validation is warranted as they were identified by the different statistical protocols.

Since Bernabeu E *et al.* focused exclusively on G×Sex interactions, it is notable that most of the overlapping signals in our analyses were primarily driven by G×Sex interactions (7/7 loci; 9/11 trait–locus pairs). On the contrary, only a subset of loci where G×Sex was identified as the top contributor in our study overlapped with Bernabeu E *et al.* (5 out of 36 loci for UKB1 and 4 out of 11 loci for BBJ1). This observation suggests that our inclusion of a broader set of environmental factors improved the power to detect G×E interactions." (**Supplementary Note 3**).

7. Provide rationale for dividing the UK Biobank by British European and non-British European participants for discovery and replication. Given the authors are already testing for cross-population replication in two different biobanks/ancestries as their next step, examining consistency and replication within the same population should also be given emphasis. Randomly allocation of participants in UKB into discovery and replication cohorts is more appropriate.

Response:

Thank you for your comment. As described in the Methods section of the initial manuscript, we used the British European subpopulation for the discovery cohort to minimize the bias due to fine-scale population stratification during variant discovery.

Genetic studies have commonly defined populations at the continental level (e.g., Wei Z *et al.* ***Cell Genom*** (2023); The All of Us Research Program Genomics Investigators, ***Nature*** (2024)). In line with this convention, the approach you kindly suggested (i.e., random allocation) ensured that the discovery and replication cohorts were matched at the continental level (i.e., European population). Importantly, **our approach also ensured population matching at the continental level.** In the revised manuscript, we evaluated the replication in the European population in an independent cohort, All of Us. Following previous studies (The All of Us Research Program Genomics Investigators, ***Nature*** (2024)), we employed the continental-level population assignments in All of Us. As the European population appeared as a continuum in the principal component spaces, it was challenging, and potentially even ethically inappropriate, to define finer-scale populations in All of Us.

Considering these points, we believe that our approach strikes a balance between controlling for population stratification in the discovery phase and evaluating replication in both intra- and inter-cohort manners. Therefore, we prefer to retain our initial approach for dividing UKB into discovery and replication cohorts.

Modification:

We added the following sentence to clarify the rationale in the Results section, in addition to the Methods section, where we initially presented the rationale mentioned above.

"In UKB, we used British European participants as a discovery cohort (UKB1, N_{max} = 273,453) and other European participants as a genetically independent replication cohort (UKB2, N_{max} = 8,149) to minimize population stratification in the discovery cohort." (**Results**, P8 line 154–156).

"For the discovery cohort (UKB1), we included British Europeans, defined as the intersection of the self-reported British participants and the genetically "Caucasian" participants, to strictly reduce the inflation of test statistics due to population stratification." (**Methods**, P38 line 967–969).

8. Given the possible heterogeneity of the discovery and replication cohorts, it is less surprising that there was lower replication of GxE interactions between UKB1 and UKB2. This should be stated as a limitation in the limitations section.

Response:

Thank you for your comment. As described in our response to Comment #7, we matched the discovery and replication cohorts at the continental population level (i.e., the European population), a commonly accepted practice in genetic studies (e.g., The All of Us Research Program Genomics Investigators, ***Nature*** (2024)).

We note that **comparing the replication rates of UKB2 and BBJ2 is not appropriate, as the number of individuals was substantially different**. In addition, the highly pleiotropic G×E interactions at the *ALDH2* locus were strong enough to be significantly replicated in the independent cohort with the modest sample size (JPHC, N_{max} = 10,904), which further increased the replication rate in BBJ2. Therefore, we do not consider the lower replication rate in UKB2 compared to BBJ2 a limitation.

As we evaluated in #5, the replication rates of G×E interactions were rather largely comparable to those of GWAS. Considering that differences in environmental distributions across cohorts can reduce the replicability of G×E interactions, and that traditional G×E studies have suffered from low replication rates, these findings suggest the high reliability of the G×E interactions identified in our manuscript.

Beyond replication, we are aware that the detection power of G×E interactions could differ across populations generally. We mentioned this in the Discussion in both the initial and revised manuscripts as follows.

Modification:

We added the comparison of replication rates between G×E interactions and GWAS, as written in #5.

In both the initial and revised manuscripts, we also mentioned the difference in the detection power of G×E interactions across populations.

"In addition to differences in allele frequencies, environmental factors and their distributions could differ across populations or their correlated factors (e.g., regions), influencing the detection power of G×E interactions." (**Discussion**, P28 line 670–673).

9. If the cross-population replication was so low, why do the BBJ significant loci consist of mostly shared loci? Does this suggest that there may be more undiscovered loci in the BBJ population? Page 13, Line 248

Response:

We appreciate your comment, which provided us with the opportunity to clarify this point. First, we would like to clarify that the conservatively estimated rate of the cross-population shared G×E interactions was 0.39 in the initial manuscript. In the revised manuscript, we followed your suggestion in #11 to include proxy variants for the previously excluded loci; the revised estimate remained 0.39. Therefore, we would like to retain our interpretation that **G×E interactions were moderately shared across populations**, encouraging future cross-population meta-analysis efforts.

Regarding the BBJ findings, we reported that 40% (6/15) of the genome-wide significant loci in BBJ1 were shared with UKB1 in both initial and revised manuscripts. This result does not suggest that "most" BBJ loci were shared but rather reinforces our message of the moderate cross-population consistency of G×E interactions.

Nevertheless, we agree with you that more undiscovered loci likely remain undiscovered. Indeed, one of the key messages from the section on the cross-population consistency of G×E interactions was to encourage future biobank collaborative efforts to detect more G×E

loci. To further support this, we leveraged the large sample sizes of our analyses and **newly performed a subsampling analysis, observing that the number of discovered G×E interactions has not been saturated at the current sample sizes of UKB1 and BBJ1** (**Ext. Data Fig. 5i**). At the same time, the subsampling analysis indicated that **biobank-scale sample sizes were required to detect G×E interactions**. This provided an additional rationale for conducting the discovery stage in the European and East Asian populations, for which large-scale genome data with rich phenotypes and environmental information were available.

In summary, **our results demonstrated the moderate cross-population consistency of G×E interactions and the need for future cross-population meta-analysis efforts**. This finding is consistent with the current consensus for GWAS, for which both cross-population consistent and population-specific associations have been reported, and biobank collaborative efforts are ongoing to expand the catalog of genetic associations.

Modification:

We added the results of the subsampling analysis to the **Results** section as described in #1.

10. For replication cohorts, authors state that replicated GxE interactions were driven by the same environments as the discovery cohorts (Page 13, Line 240) but do not state this in the cross-population analysis (Lines 253 – 254). Please state whether this is true.

Response:

Thank you for your comment. Yes, we confirmed that the same environmental factors contributed to the cross-population shared G×E interactions. Additionally, we confirmed that the effect sizes were in the same direction. Importantly, we confirmed these points throughout the manuscript. We revised the manuscript to explicitly state this point to further ensure clarity for readers.

Modification:

According to your kind suggestion, we added the explicit statement as follows.

"Throughout the manuscript, we stringently verified that all replicated G×E interactions were driven by the same environments as the discovery cohorts and had consistent effect directions. Interactions failing these criteria were considered non-replicated, regardless of *P*-values." (**Results**, P10–11 line 227–230).

11. Page 14 262 – 263 Authors state that lead variants were not testable in the other population for 19 loci. Why were proxies not used for these loci? If not possible please provide justification.

Response:

Thank you for your thoughtful suggestion. In the initial manuscript, we excluded these loci because no genome-wide significant variants within these loci were genotyped or imputed in

the other biobank. We had originally opted not to use proxy variants, based on concerns that the power to evaluate the replication might be insufficient or that the signals might be driven by population-specific causal variants.

Following your suggestion, we used proxy variants for all loci except the *ALDH2* locus. The lead variant of the *ALDH2* locus (rs671) is a well-established East Asian-specific functional variant. Given its highly pleiotropic G×E interactions, we excluded this locus to minimize potential bias in estimating cross-population consistency of the G×E interactions.

Modification:

We revised the replication rate as follows. We also re-estimated the conservatively estimated rate of cross-population shared G×E interactions, as described in #9, which yielded the same value of 0.39 as reported in the initial manuscript.

"We excluded the *ALDH2* locus (rs671) due to its East Asian specificity and well-known function to avoid introducing bias in the estimate of cross-population consistency. Out of 71 trait–locus pairs, 22 were nominally, and 9 were significantly cross-population replicated ($P_{G\times E} < 0.05/71 = 7.0\times10^{-4}$). The conservative signal-sharing estimate (Storey's π_1) was 0.39, indicating moderate consistency of G×E interactions across populations." (**Results**, P12 line 274–279).

12. Line 514 – 515, does this imply that PGS should be applied in consistent environmental settings?

Response:

Yes. Beyond the environment-stratified PGS, **we newly investigated the utility of PGS constructed from G×E interactions (G×E-PGS)** for phenotype prediction in the revised manuscript. We first applied the G×E polygenic score (G×Eprs) method (Jayasinghe D *et al.*, ***Genet Epidemiol*** (2024)), as introduced by Reviewer #1. This approach incorporates a PGS constructed from G×E interactions (G×E-PGS) into a multiple regression model, alongside the additive-effect PGS, and tests whether the G×E-PGS has a significant non-zero effect size. As G×E-aware PGS construction methods have not been established, we constructed all types of PGS using PRScs-auto, one of the most widely used tools for PGS construction. Tested in independent cohorts (the European population in All of Us and the East Asian population in J-MICC/HERPACC), **G×E-PGS significantly explained phenotypic variance for 11/23 and 9/23 trait–environment pairs within and across populations, respectively**. These results supported the potential of G×E-PGS for contributing to phenotype prediction.

Notably, **the G×E-PGS constructed from G×Sex interactions stratified BMI in opposite directions between males and females in J-MICC/HERPACC** (**Figure 4h**). This stratification was apparent even within each sex-stratified group and among individuals with similar marginal PGS, **suggesting the potential for extending PGS into two dimensions to improve phenotype prediction**. Recent studies have increasingly focused on the interactions between marginal-effect PGS and environments (often referred to as PGS×E) (Boye C *et al.* ***Nat Genet*** (2024)). However, the heterogeneity observed within individuals of the same sex and with similar marginal-effect PGS cannot be captured by PGS×E, indicating that G×E-PGS more effectively reflects the polygenic architecture underlying sex differences.

A prediction model incorporating G×E-PGS, additive- and marginal-effect PGS, and sex showed a 16% increase in prediction accuracy for BMI compared to the model without G×E-PGS (R^2 = 0.128 versus 0.110, 10-fold cross-validation in J-MICC/HERPACC). These findings demonstrate the utility of G×E-PGS in improving BMI prediction by capturing interaction effects missed by conventional PGS. On the contrary, the gain in prediction accuracy was modest for other trait–environment pairs (**Supplementary Table 25**). As we used biobank-scale large sample sizes and constructed PGS using one of the state-of-the-art methods, these results may show the current challenges in the G×E-PGS framework in general. Larger sample sizes and/or methods tailored for G×E-aware PGS construction are warranted to fully realize the potential of G×E interactions in precision medicine.

According to your thoughtful comment, we added the prediction analysis using G×E-PGS.

"Consistent with these findings, PGS constructed from G×E interactions (G×E-PGS) significantly explained phenotypic variance in independent cohorts (11/23 and 9/23 trait–environment pairs within and across populations, respectively; **Supplementary Methods and Supplementary Table 25**). Notably, the G×Sex-based PGS stratified BMI in opposite directions between males and females in J-MICC/HERPACC, capturing the polygenic architecture of sex differences (**Figure 4h**). This stratification was apparent even within each sex and among individuals with similar marginal PGS, suggesting that PGS extended to two dimensions could enhance phenotype prediction. Indeed, a model incorporating G×E-PGS, additive- and marginal-effect PGS, and sex improved BMI prediction accuracy by 16% over a model without G×E-PGS (R^2 = 0.128 versus 0.110, 10-fold cross-validation in J-MICC/HERPACC). In contrast, gains in prediction accuracy were modest for other trait–environment pairs (**Supplementary Table 25**). Larger sample sizes and methods tailored for G×E-PGS construction are warranted to fully realize the potential of G×E interactions in precision medicine." (**Results**, P20–21 line 485–498).

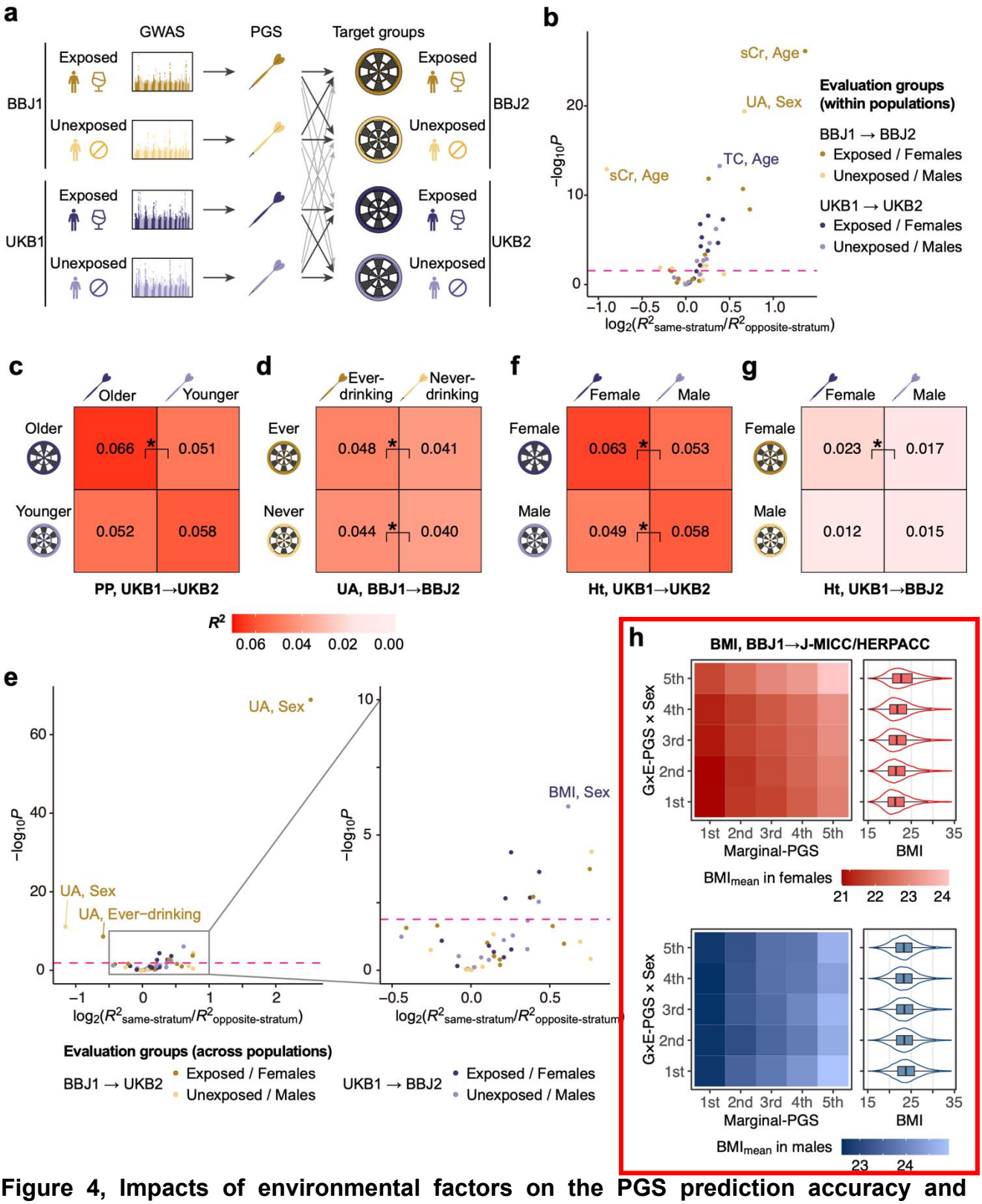


Figure 4, Impacts of environmental factors on the PGS prediction accuracy and portability

a, overview of polygenic score (PGS) construction and evaluation of within-population prediction accuracy and cross-population portability. Wine icons from Phosphor (https://phosphoricons.com/) under MIT license. **b**, PGS prediction accuracy in the same population. *X*-axis: $\log_2$ ratio of prediction accuracy (R^2) between "same-stratum" and "opposite-stratum" PGS (based on matching vs. non-matching environmental strata). *Y*-axis:

P-values of R^2 differences. Dashed line: FDR threshold of 0.05. **c** and **d**, examples of within-population prediction accuracy (R^2): pulse pressure (PP) in UKB stratified by age (**c**) and uric acid (UA) in BBJ stratified by drinking status (**d**). Asterisks: FDR < 0.05. Figure designs inspired by the referenced studies. **e**, same as **b**, for cross-population portability. **f** and **g**, examples of PGS prediction accuracy within (**f**) and across (**g**) populations, for hematocrit (Ht) stratified by sex. **h**, BMI distribution stratified by PGS constructed from G×E interactions (G×E-PGS) and that from marginal PGS. Boxplots show median, quartiles, and 1.5× interquartile range whiskers.

"**Phenotype prediction using PGS constructed from G×E interactions (G×E-PGS)**

The G×Eprs framework

We employed the G×Eprs framework to test whether G×E-PGS significantly explains the phenotypic variation. This framework fits the following linear regression model:

$$Y \sim PGS_{additive} + PGS_{G\times E} \times E + PGS_{G\times E} + E$$

where $PGS_{additive}$ is the PGS constructed from additive genetic effects, $PGS_{G\times E}$ is PGS constructed from G×E interaction effects, and E denotes the environmental factor. We also refer to the PGS constructed from marginal effects as $PGS_{marginal}$ below. As G×E-aware PGS construction methods have not been established, we constructed all types of PGS using PRScs-auto with the 1000 Genome Project reference panel for the corresponding population. For the response variable Y, we used phenotype residuals obtained by regressing the raw phenotype values on the same covariates used in the G×E interaction tests. To preserve the variance attributable to environmental factors, we excluded age×sex, age^2, and age^2×sex from the covariates when testing G×Age PGS and excluded age×sex and age^2×sex when testing G×Sex PGS. Following the original study, we evaluated the significance of the $PGS_{G\times E} \times E$ term using both Wald *P*-values and permutation-based *P*-values, and considered it significant when FDR was less than 0.05 for both *P*-values. Permutation tests were performed with 10,000 iterations by randomly shuffling the values of $PGS_{G\times E} \times E$, as implemented in the R package "GxEprs".

Phenotype prediction using G×E-PGS

We compared two elastic net models for phenotype prediction, one with G×E-PGS and one without G×E-PGS. The baseline model included marginal PGS and the environmental factor (E), assuming a standard PGS constructed based on GWAS summary statistics:

$$Y \sim PGS_{marginal} + E$$

The full model additionally included additive PGS, G×E-PGS, and the interaction with the environment:

$$Y \sim PGS_{marginal} + PGS_{additive} + PGS_{G\times E} + PGS_{G\times E} \times E + E$$

We also included Age^2 into both models when testing G×Age-PGS to account for non-linear variation in phenotype values across age. Models were implemented using the R package "glmnet" (version 4.1.8), which performs regularized regression by linearly combining L1 (Lasso) and L2 (Ridge) penalties. The hyperparameters (Lasso and Ridge penalties) were tuned via a grid search over a range from 0 to 1 in increments of 0.1, and were selected

based on 10-fold cross-validation by minimizing the mean squared error on the validation folds. We reported the R^2 from the 10-fold cross-validation as the prediction accuracy." (**Supplementary Methods**).

13. Could the effect of drinking and the ALDH2 locus on T2D be mediated through hemoglobin levels, and be unrelated to metabolic disease?

Response:

Thank you for your comment. The G×E interaction for T2D remained significant when conditioning on the hemoglobin measurements ($P_{G\times E} = 8.1\times10^{-15}$), which was comparable to the unconditioned *P*-value ($P_{G\times E} = 4.8\times10^{-16}$). To ensure a fully transparent comparison, we re-estimated the unconditioned *P*-value using only participants with available hemoglobin measurements, and the *P*-value was even closer to the conditioned one ($P_{G\times E} = 2.1\times10^{-15}$). These results suggest that hemoglobin plays only a minimal mediating role in the G×E interaction for T2D.

Modification:

We added the conditioning analysis in the **Results**.

"These included a G×E-specific signal at the *ALDH2* locus for hemoglobin (Hb) in BBJ1 ($P_{G\times E} = 2.2\times10^{-15}$ and $P_{marginal} = 1.7\times10^{-3}$, primarily driven by G×Ever-drinking), which was replicated in BBJ2 ($P_{G\times E} = 2.9\times10^{-9}$ and $P_{marginal} = 8.9\times10^{-3}$). This signal demonstrated the power of G×E interactions to detect context-specific genetic effects. The same locus also exhibited a significantly replicated G×Ever-drinking interaction for type 2 diabetes (T2D) ($P_{G\times E} = 4.8\times10^{-16}$ and 1.2×10^{-6} in BBJ1 and BBJ2, respectively). This G×E interaction remained significant after adjusting for Hb ($P_{G\times E} = 8.1\times10^{-15}$ in BBJ1), suggesting minimal mediation by Hb." (**Results**, P11 line 235–242).

14. Do you have interaction effects for natto x PITX2 locus on arrhthmia

Response:

Thank you for your comment. Yes, we have reported the interaction effects between natto consumption and the *PITX2* locus on arrhythmia. We observed that G×Sex, G×Ever-drinking, and G×Ever-smoking also contributed to this locus. Nevertheless, **testing the G×Natto interaction alone surpassed the genome-wide significant threshold ($P_{G\times E} = 2.1\times10^{-10}$)**, supporting the primary contribution of G×Natto interaction at this locus. To further validate this G×E interaction, **we newly compared the frequency of natto consumption in the same individuals before and after initiating warfarin therapy. We observed a marked decrease in natto consumption after starting warfarin medication** (**Figure 2i**), further supporting the reverse causal direction from warfarin medication to the decreased natto consumption. We also conducted a Mendelian randomization analysis, which revealed a significant causal association from warfarin medication to decreased natto consumption ($P = 6.6\times10^{-3}$). We performed a Cox regression analysis using incidental cases with disease

onsets at least six months after baseline. The G×E interaction was not significant at this locus in BBJ1 ($P_{G\times E}$ = 0.077), which was expected given that the Cox regression model tested the temporal direction from environments to diseases, opposite to the temporal order observed above. Together, these results support the reverse causal relationship from the disease to the environment, consistent with clinical practice in Japan, where physicians commonly advise warfarin-taking patients to avoid eating natto.

Following your insightful comments here and in Comment #15, we further explored the potential mechanism of this G×E interaction. **We found that the marginal effect size of the lead variant was heterogeneous between the subgroups of arrhythmia**, making the overall effect size for arrhythmia dependent on the rate of atrial fibrillation/flutter patients among the arrhythmia patients. This led to a difference in the effect size for arrhythmia across strata of natto consumption frequency, **which explained the link between the reverse causality and the G×Natto interaction**. In line with this notion, the G×Natto interaction was not significant when evaluated separately within the atrial fibrillation/flutter and the other arrhythmia subgroups ($P_{G\times E}$ = 0.84 and 2.6×10^{-4}, respectively).

Modification:

We added the additional evidence to support the reverse causal direction in **Results**.

"Testing G×Natto alone surpassed genome-wide significance ($P_{G\times E} = 2.1\times10^{-10}$), supporting its primary role at this locus." (**Results**, P15 line 350–351).

"Indeed, natto intake declined markedly after warfarin initiation in the same individuals (**Figure 2i**). Mendelian randomization supported the negative causal effect of warfarin on natto consumption ($P = 6.6\times10^{-3}$, effect size = -0.029 [95% confidence interval (CI), -0.051 to -8.2×10^{-3}]).

When stratifying arrhythmia patients by clinical subgroups, the marginal effect size of rs72900155 was substantially larger for atrial fibrillation/flutter than for other subgroups (0.49 [95% CI, 0.46–0.52] versus 0.13 [0.10–0.15]). Due to this effect size heterogeneity, the overall effect size for arrhythmia would depend on the proportion of atrial fibrillation/flutter patients and therefore differ across natto intake strata, explaining the link between the reverse causality and the G×Natto interaction. Consistently, the G×Natto interaction was not significant within either subgroup when evaluated separately ($P_{G\times E}$ = 0.84 for atrial fibrillation/flutter; 2.6×10^{-4} for other subgroups)." (**Results**, P16 line 360–371).

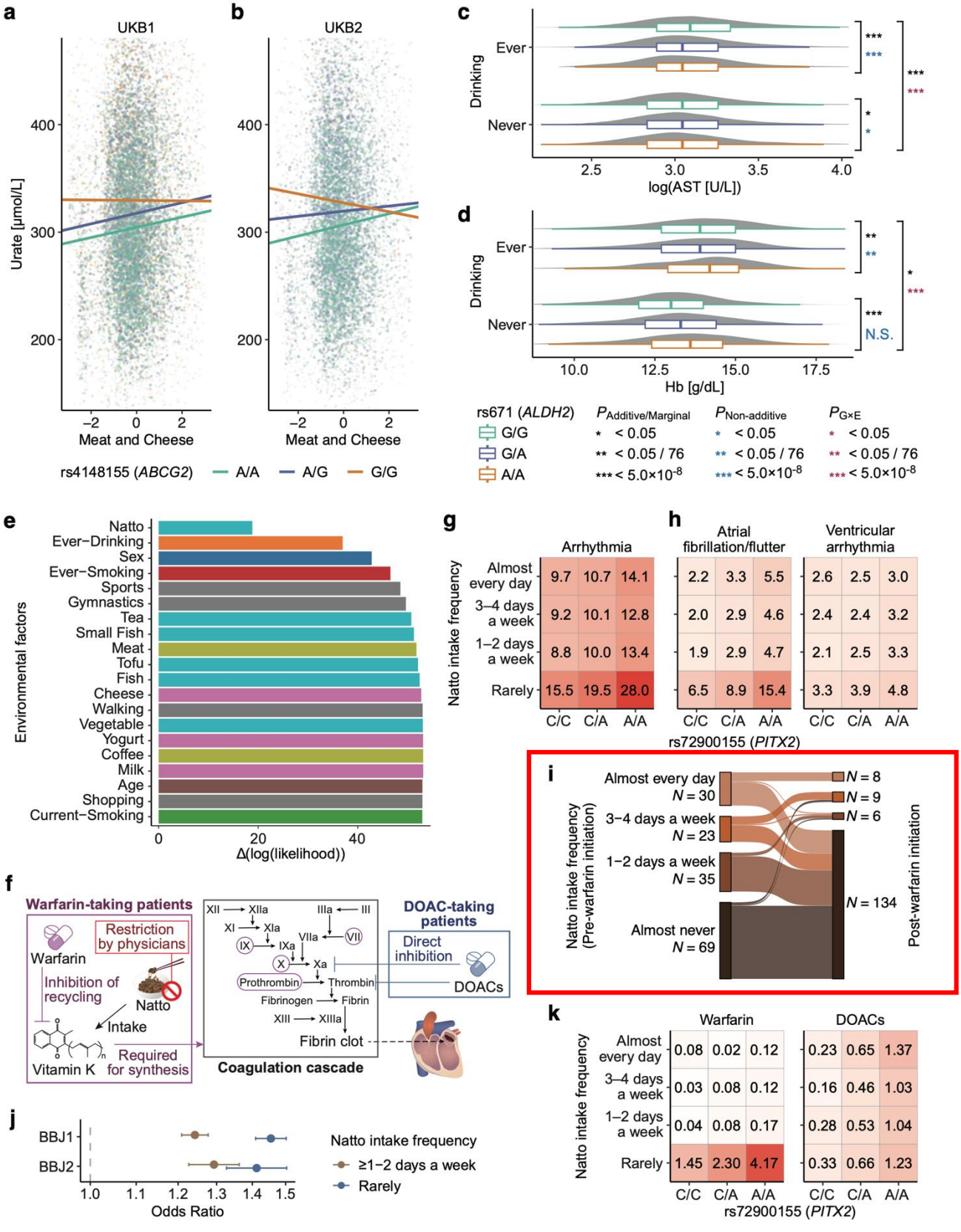


Figure 2, Representative loci with G×E interactions.

a and **b**, Urate levels across rs4148155 genotypes (the *ABCG2* locus) and "meat and cheese" consumption in UKB1 (**a**) and UKB2 (**b**). Dots show 5,000 randomly sampled individual measurements per genotype; lines represent the linear regression lines using all participants. **c** and **d**, AST (**c**) and hemoglobin (Hb, **d**) levels by rs671 genotypes (the *ALDH2* locus) and drinking status in BBJ1. Boxplots show median, quartiles, and 1.5× interquartile range whiskers. **e**, Log-likelihood improvement for G×E models of rs72900155 (the *PITX2* locus) for arrhythmia. Environmental factors were added stepwise to the null model (**Methods**). **f**, schematic of the mechanisms of action for warfarin, natto (fermented soybean), and direct oral anticoagulants (DOAC). Heart icon from TogoTV (© 2016 DBCLS TogoTV, CC-BY-4.0). **g**, heatmap of the arrhythmia prevalence in BBJ1. **h**, same as **g**, stratified by arrhythmia subgroups. **i**, natto intake before and after warfarin initiation in BBJ atrial fibrillation/flutter patients with ≥1 PT-INR record per year, considering that regular monitoring of PT-INR is required during warfarin therapy. **j**, odds ratios of rs72900155 (the *PITX2* locus) for arrhythmia stratified the natto intake frequency. Error bars represent 95% confidence intervals (CI; Firth logistic regression). **k**, heatmap of the atrial fibrillation/flutter prevalence in BBJ2, stratified by warfarin and DOAC medication.

We also described the Cox regression analysis using incidental cases in **Supplementary Note 8**.

"On the other hand, the G×E interaction for arrhythmia at the *PITX2* locus was not significant ($P_{G\times E}$ = 0.077). In the section "A disease-induced dietary change caused a G×E interaction," we discussed that this locus was driven by a reverse causal relationship from a disease to an environment (specifically, from warfarin medication for atrial fibrillation/flutter to the avoidance of natto consumption). Considering that the Cox regression model tested the temporal direction from environments to diseases, the failure of its verification in the Cox regression was expected." (**Supplementary Note 8**).

"All statistical tests were two-sided unless otherwise noted. As phenotypes, we targeted quantitative traits (clinical biomarkers, metabolites, and protein expression measurements), dichotomous traits (disease statuses), and time-to-event traits (disease onset for incidental cases and overall survival). For quantitative and dichotomous traits, we tested G×E interactions using GEM, a fast and scalable implementation of linear and logistic regressions with G×E interaction terms[23] (**Ext. Data Fig. 1**). We estimated the model-robust "sandwich" standard errors to suppress the inflation of statistics[23]. As the case–control imbalance can cause the inflation of test statistics in the logistic regression, we re-estimated the effect sizes and *P*-values using Firth logistic regression for the variants with *P*-values less than 5.0×10^{-8} for dichotomous traits. For the time-to-event traits, we employed the Cox proportional hazard model implemented in the R package "survival". Individuals who had already developed the target disease at baseline or developed it within six months from baseline were excluded from the incidental case analyses." (**Methods**, P43 line 1077–1089).

15. Line 375, Can you include a significance test demonstrating that the proportion of arrhythmia patients is higher in the homozygous carriers of natto non-consumers? I think figure 2g,h, and i are misleading without this test for comparison. While it does appear that among AA carriers there is a higher rate of individuals that do not eat natto than AC or CC

genotypes, there is no difference in the proportion of individuals who do not eat natto in the other plots. The bottom right corner of these figures appears darker simply because that genotype and not eating natto are more common. I am therefore skeptical or unclear about the warfarin findings. Please explain more clearly with statistical tests.

Response:

Thank you for your thoughtful comment, which inspired us to investigate the potential statistical mechanism of this G×E interaction. As discussed in our response to Comment #14, **we newly proposed the potential statistical mechanism that natto consumption caused this G×E interaction through heterogeneous effect sizes of the *PITX2* locus across arrhythmia subgroups.** Therefore, the proportion of arrhythmia patients was not necessarily higher in the homozygous carriers of natto non-consumers when tested in individual arrhythmia subgroups. We indeed confirmed that the G×Natto interaction was genome-wide significant for all arrhythmia ($P_{G\times E} = 2.1\times10^{-10}$) but not when evaluated in individual arrhythmia subgroups ($P_{G\times E}$ = 0.84 and 2.6×10^{-4} for atrial fibrillation/flutter and the other arrhythmia subgroups, respectively).

Modification:

We discussed the potential statistical mechanism of this G×E interaction in **Results** as written in #14, which included the significance tests for the G×Natto interaction for all arrhythmia and for the subgroups of arrhythmia.

16. Please discuss the possibility of survival bias resulting in age bins that are heterogeneous, therefore explaining some of the age-interaction effects.

Response:

Thank you for your thoughtful comment. We agree with you that survival bias may influence G×E interactions. We have now explicitly discussed the potential impact of survival bias on G×Age interactions in the **Discussion** section of the revised manuscript. We observed a significant G×E interaction on overall survival at the *ALDH2* locus in BBJ1, driven by interactions with sex, ever-drinking, and age ($P = 1.7\times10^{-11}$). This result suggests that the G×E interaction may also affect human lifespan.

Modification:

We mentioned the potential survival bias for G×Age interactions in **Discussion**.

“We also note that environment-specific bias, such as survival bias for G×Age interactions, requires careful consideration.” (**Discussion**, P27 line 650–651).

We also added the G×E interaction on overall survival in **Results and Supplementary Note 8**.

“We also performed Cox analyses for overall survival, where the temporal order from environments to the endpoint of death was clear, and observed a significant G×E interaction

at the *ALDH2* locus driven by interactions with sex, ever-drinking, and age ($P = 1.7\times10^{-11}$), indicating that the G×E interaction may affect human lifespan." (**Results**, P17 line 387–390).

"To assess the causal direction more confidently, we performed Cox regression for overall survival, where the temporal order from environmental factors to the endpoint of death was apparent. Testing 53 loci in their discovered cohorts, we observed a significant G×E interaction for the *ALDH2* locus in BBJ1 ($P_{G\times E} = 1.7\times10^{-11}$; driven by G×Sex, G×Ever-drinking, and G×Age), indicating that the G×E interaction may affect human lifespan. In UKB1, we observed a significant G×Age interaction at the *APOE* locus ($P_{G\times E} = 5.5\times10^{-8}$), consistent with its G×Age interactions for total cholesterol (TC) and triglycerides (TG). As its effect size was in the same direction as the additive genetic effect (0.098 [95% confidence interval, 0.065–0.130] and 0.054 [0.020–0.087], respectively), this interaction suggested a larger mortality risk for carriers of the risk variant than non-carriers at older ages." (**Supplementary Note 8**).

Reviewers #4:

Understanding how genetic variant effects on health outcomes differ by environmental contexts is an important and timely area of research which can address emerging needs and gaps in the field. An incredibly large number of analyses and enormous scope of work was conducted, using 2 large biobank samples consisting of over 500,000 participants, to generate a resource of GxE interactions for the scientific community and provide exemplar GxE insights into biologic mechanisms for health outcomes.

The authors performed genome-wide gene-environment interaction (GxE) analyses for 47 phenotypes and 12 environments, identifying 92 GxE at 53 loci (several of which have been previously reported) in total. Additional GxE analyses were performed for just under 3,000 protein and 325 metabolite measures. Interestingly, they report molecular changes (metabolome/proteome) related to loci identified in the GxE scan, suggesting biologic pathways involved in gene-environment interplay. Finally, pleiotropic G×E interactions were tested for another 87 diseases across the 2 biobanks, GxE heritability and genetic correlation and differences in polygenic prediction accuracy in different environmental contexts was assessed. The authors report that most of the pleiotropic GxE interactions occurred within a set of related traits and were not commonly observed across phenotypic trait categories. Single cell analysis revealed cell types and biology relevant to the observed differences in genetic effects on pulse pressure by age strata.

Overall, this work is highly significant and I found the methods employed to be rigorous and thorough. Additional strengths of the paper include discovery and replication samples and across 2 populations, “re”discovery of well known GxE and identification of new GxE and considerable secondary analyses to highlight the biologic relevance of these findings. I do have some concerns below that should be addressed.

Response:

We sincerely thank you for taking your time to consider our manuscript and for kindly recognizing its value. We truly appreciate your insightful comments, which helped us further improve the manuscript. To improve readability, the changes made to the figures are highlighted with red boxes.

1-1) The shaping of the introduction and utility of GxE results for precision medicine purposes is important and makes sense. What I struggle with a bit is that all of the analyses and plots and interpretations we can make are at the population level. There seems to be considerable heterogeneity at the individual participant level – how informative will the results really be for precision medicine at an individual level? Can these results really predict on an individual level?

Response:

Thank you for your thoughtful comment. We fully agree with you that bridging population-level discoveries to individual-level utility is a key challenge in translating G×E interaction

findings into precision medicine. In our manuscript, we reported that G×E heritability was significantly positive for 18 biobank-trait pairs at the genome-wide level, with the mean G×E heritability of 3.26%. The heritability determines the upper bound for the prediction accuracy that an ideal polygenic score (PGS) could theoretically achieve (O'Conner LJ. ***Nat Genet*** (2021)). Therefore, our results suggest that incorporating G×E interactions may have the potential to improve the prediction accuracy of PGS at a clinically meaningful level in theory. However, it is well recognized that actual prediction accuracy achieved by real-world PGS models often falls short of heritability estimates, even for marginal genetic effects (O'Conner LJ *et al.* ***Nat Genet*** (2021)). Whether the clinically meaningful gain for PGS prediction can be realized in practice from G×E interactions remains an open question.

To address this, **we newly investigated the utility of PGS constructed from G×E interactions (G×E-PGS)** for phenotype prediction in the revised manuscript. As G×E-aware PGS construction methods have not been established, we constructed all types of PGS using PRScs-auto, one of the most widely used tools for PGS construction. We first applied the G×E polygenic score (G×Eprs) method (Jayasinghe D *et al.*, ***Genet Epidemiol*** (2024)), as introduced by Reviewer #1. This approach incorporates a PGS constructed from G×E interactions (G×E-PGS) into a multiple regression model, alongside the additive-effect PGS, and tests whether the G×E-PGS has a significant non-zero effect size. Tested in independent cohorts (the European population in All of Us and the East Asian population in J-MICC/HERPACC), **G×E-PGS significantly explained phenotypic variance for 11/23 and 9/23 trait–environment pairs within and across populations, respectively**. These results supported the potential of G×E-PGS for contributing to phenotype prediction.

Notably, **the G×E-PGS constructed from G×Sex interactions stratified BMI in opposite directions between males and females in J-MICC/HERPACC** (**Figure 4h**). This stratification was apparent even within each sex-stratified group and among individuals with similar marginal PGS, **suggesting the potential of extending PGS into two dimensions to improve phenotype prediction**. Recent studies have increasingly focused on the interactions between marginal-effect PGS and environments (often referred to as PGS×E) (Boye C *et al.* ***Nat Genet*** (2024)). However, the heterogeneity observed within individuals of the same sex and with similar marginal-effect PGS cannot be captured by PGS×E, indicating that G×E-PGS more effectively reflects the polygenic architecture underlying sex differences.

A prediction model incorporating G×E-PGS, additive- and marginal-effect PGS, and sex showed a 16% increase in prediction accuracy for BMI compared to the model without G×E-PGS (R^2 = 0.128 versus 0.110, 10-fold cross-validation in J-MICC/HERPACC). These findings demonstrate the utility of G×E-PGS in improving BMI prediction by capturing interaction effects missed by conventional PGS. On the contrary, the gain in prediction accuracy was modest for other trait–environment pairs (**Supplementary Table 25**). As we used biobank-scale large sample sizes and constructed PGS using one of the state-of-the-art methods, these results may show the current challenges in the G×E-PGS framework in general. Larger sample sizes and/or methods tailored for G×E-aware PGS construction are warranted to fully realize the potential of G×E interactions in precision medicine.

Modification:

According to your thoughtful comment, we added the prediction analysis using G×E-PGS.

"Consistent with these findings, PGS constructed from G×E interactions (G×E-PGS) significantly explained phenotypic variance in independent cohorts (11/23 and 9/23 trait–environment pairs within and across populations, respectively; **Supplementary Methods**

and Supplementary Table 25). Notably, the G×Sex-based PGS stratified BMI in opposite directions between males and females in J-MICC/HERPACC, capturing the polygenic architecture of sex differences (**Figure 4h**). This stratification was apparent even within each sex and among individuals with similar marginal PGS, suggesting that PGS extended to two dimensions could enhance phenotype prediction. Indeed, a model incorporating G×E-PGS, additive- and marginal-effect PGS, and sex improved BMI prediction accuracy by 16% over a model without G×E-PGS (R^2 = 0.128 versus 0.110, 10-fold cross-validation in J-MICC/HERPACC). In contrast, gains in prediction accuracy were modest for other trait–environment pairs (**Supplementary Table 25**). Larger sample sizes and methods tailored for G×E-PGS construction are warranted to fully realize the potential of G×E interactions in precision medicine." (**Results**, P20–21 line 485–498).

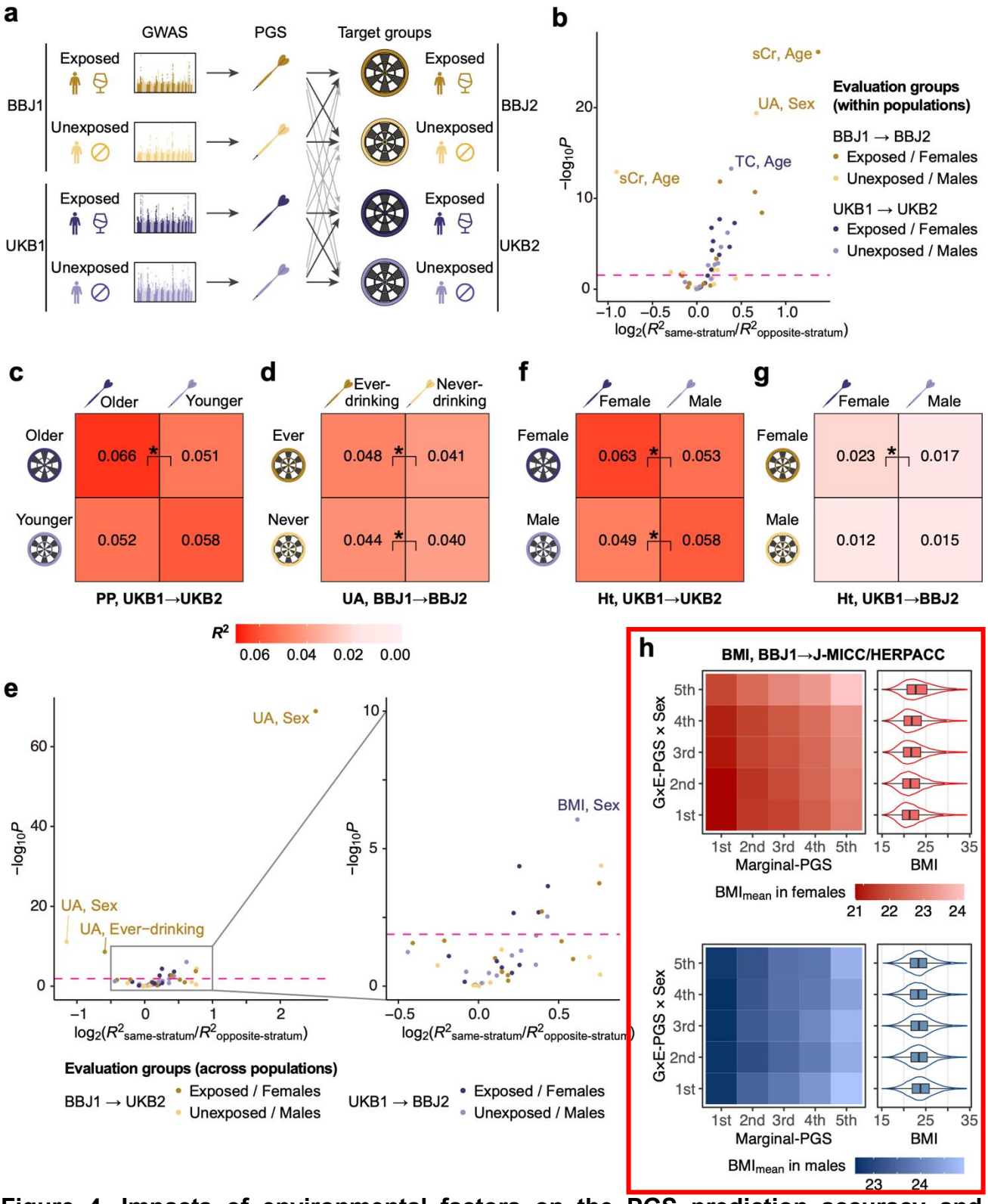


Figure 4, Impacts of environmental factors on the PGS prediction accuracy and portability

a, overview of polygenic score (PGS) construction and evaluation of within-population prediction accuracy and cross-population portability. Wine icons from Phosphor (https://phosphoricons.com/) under MIT license. **b**, PGS prediction accuracy in the same population. *X*-axis: $\log_2$ ratio of prediction accuracy (R^2) between "same-stratum" and "opposite-stratum" PGS (based on matching vs. non-matching environmental strata). *Y*-axis:

P-values of R^2 differences. Dashed line: FDR threshold of 0.05. **c** and **d**, examples of within-population prediction accuracy (R^2): pulse pressure (PP) in UKB stratified by age (**c**) and uric acid (UA) in BBJ stratified by drinking status (**d**). Asterisks: FDR < 0.05. Figure designs inspired by the referenced studies. **e**, same as **b**, for cross-population portability. **f** and **g**, examples of PGS prediction accuracy within (**f**) and across (**g**) populations, for hematocrit (Ht) stratified by sex. **h**, BMI distribution stratified by PGS constructed from G×E interactions (G×E-PGS) and that from marginal PGS. Boxplots show median, quartiles, and 1.5× interquartile range whiskers.

"**Phenotype prediction using PGS constructed from G×E interactions (G×E-PGS)**

The G×Eprs framework

We employed the G×Eprs framework to test whether G×E-PGS significantly explains the phenotypic variation. This framework fits the following linear regression model:

$$Y \sim PGS_{additive} + PGS_{G\times E} \times E + PGS_{G\times E} + E$$

where $PGS_{additive}$ is the PGS constructed from additive genetic effects, $PGS_{G\times E}$ is PGS constructed from G×E interaction effects, and E denotes the environmental factor. We also refer to the PGS constructed from marginal effects as $PGS_{marginal}$ below. As G×E-aware PGS construction methods have not been established, we constructed all types of PGS using PRScs-auto with the 1000 Genome Project reference panel for the corresponding population. For the response variable Y, we used phenotype residuals obtained by regressing the raw phenotype values on the same covariates used in the G×E interaction tests. To preserve the variance attributable to environmental factors, we excluded age×sex, age^2, and age^2×sex from the covariates when testing G×Age PGS and excluded age×sex and age^2×sex when testing G×Sex PGS. Following the original study, we evaluated the significance of the $PGS_{G\times E} \times E$ term using both Wald *P*-values and permutation-based *P*-values, and considered it significant when FDR was less than 0.05 for both *P*-values. Permutation tests were performed with 10,000 iterations by randomly shuffling the values of $PGS_{G\times E} \times E$, as implemented in the R package "GxEprs".

Phenotype prediction using G×E-PGS

We compared two elastic net models for phenotype prediction, one with G×E-PGS and one without G×E-PGS. The baseline model included marginal PGS and the environmental factor (E), assuming a standard PGS constructed based on GWAS summary statistics:

$$Y \sim PGS_{marginal} + E$$

The full model additionally included additive PGS, G×E-PGS, and the interaction with the environment:

$$Y \sim PGS_{marginal} + PGS_{additive} + PGS_{G\times E} + PGS_{G\times E} \times E + E$$

We also included Age^2 into both models when testing G×Age-PGS to account for non-linear variation in phenotype values across age. Models were implemented using the R package "glmnet" (version 4.1.8), which performs regularized regression by linearly combining L1 (Lasso) and L2 (Ridge) penalties. The hyperparameters (Lasso and Ridge penalties) were tuned via a grid search over a range from 0 to 1 in increments of 0.1, and were selected based on 10-fold cross-validation by minimizing the mean squared error on the validation

folds. We reported the R^2 from the 10-fold cross-validation as the prediction accuracy." (**Supplementary Methods**).

1-2) The UKBB is not representative of the population, given it is largely comprised of healthy volunteers, which has been shown to influence GWAS estimates: https://doi.org/10.1038/s41562-023-01579-9. <https://doi.org/10.1038/s41562-023-01579-9.> This should be recognized in the discussion including whether the authors think this is likely to influence their results.

Response:

Thank you for your helpful comment. We fully agree with you that UKB is not representative of the population, as the participants were known to be healthier than the general population. To evaluate the robustness of the G×E interactions beyond UKB, **we tested for replication in the newly added four independent cohorts with diverse populations (All of Us, J-MICC/HERPACC, JPHC, and HPP; comprising European, East Asian, African, American, and Israeli populations; *N* = 436,272)**

First, we conducted within-population replication analyses in independent cohorts: the European population in All of Us (N_{max} = 208,700; in the US) and the East Asian population in J-MICC/HERPACC (N_{max} = 70,909; in Japan) for replicating the UKB1 and BBJ1 results, respectively. In these cohorts, 27% (17/62) and 56% (20/36) trait–locus pairs were significantly replicated ($P_{G\times E} < 0.05/(62+36) = 5.1\times10^{-4}$), respectively. Notably, the pleiotropic G×E interactions at the *ALDH2* locus showed a particularly high replication rate in J-MICC/HERPACC (81% = 17/21). We also tested replication in another Japanese cohort, JPHC (N_{max} = 10,904), where 6 G×E interactions at the *ALDH2* locus were again significantly replicated.

We also expanded the cross-population replication analyses to include the African and American populations in All of Us (N_{max} = 70,558 and 66,556, respectively) and the Israeli population in HPP (N_{max} = 8,645, largely composed of Ashkenazi Jewish people). In the African population, three G×E interactions were significantly replicated, and two of them were included in the shared G×E interactions between UKB1 and BBJ1. Two G×E interactions for pulse pressure were significantly replicated in the American population, both of which were originally detected in UKB1 and primarily driven by G×Age in both UKB1 and All of Us. G×E interactions were not significantly replicated in the Israeli population, possibly due to its relatively small sample size. Nevertheless, we observed the top signal for one of the shared G×E interactions between UKB1 and BBJ1 (*UGT1A* gene family for total bilirubin, commonly driven by G×Current-smoking and G×Age), suggesting its consistency across diverse populations. We summarized the replication statuses, including nominal replication with $P_{G\times E}$ less than 0.05, in **Ext. Data Fig. 5a and b**. We note that we stringently verified that all replicated G×E interactions were driven by the same environments as the discovery cohorts, and the effect sizes were in the same direction. We regarded any G×E interactions that did not meet these criteria as non-replicated, regardless of their *P*-values.

Among the independent cohorts, All of Us is a nationwide cohort, and J-MICC/HERPACC and JPHC are population-based cohorts; the participants of HPP were recruited from volunteers in Israel. **As these cohorts represent diverse populations with different recruitment strategies, questionnaires, and healthcare systems, replication in them supports the generalizability of our findings across heterogeneous settings**.

Modification:

We have included the replication in the completely independent cohorts in the **Results** section.

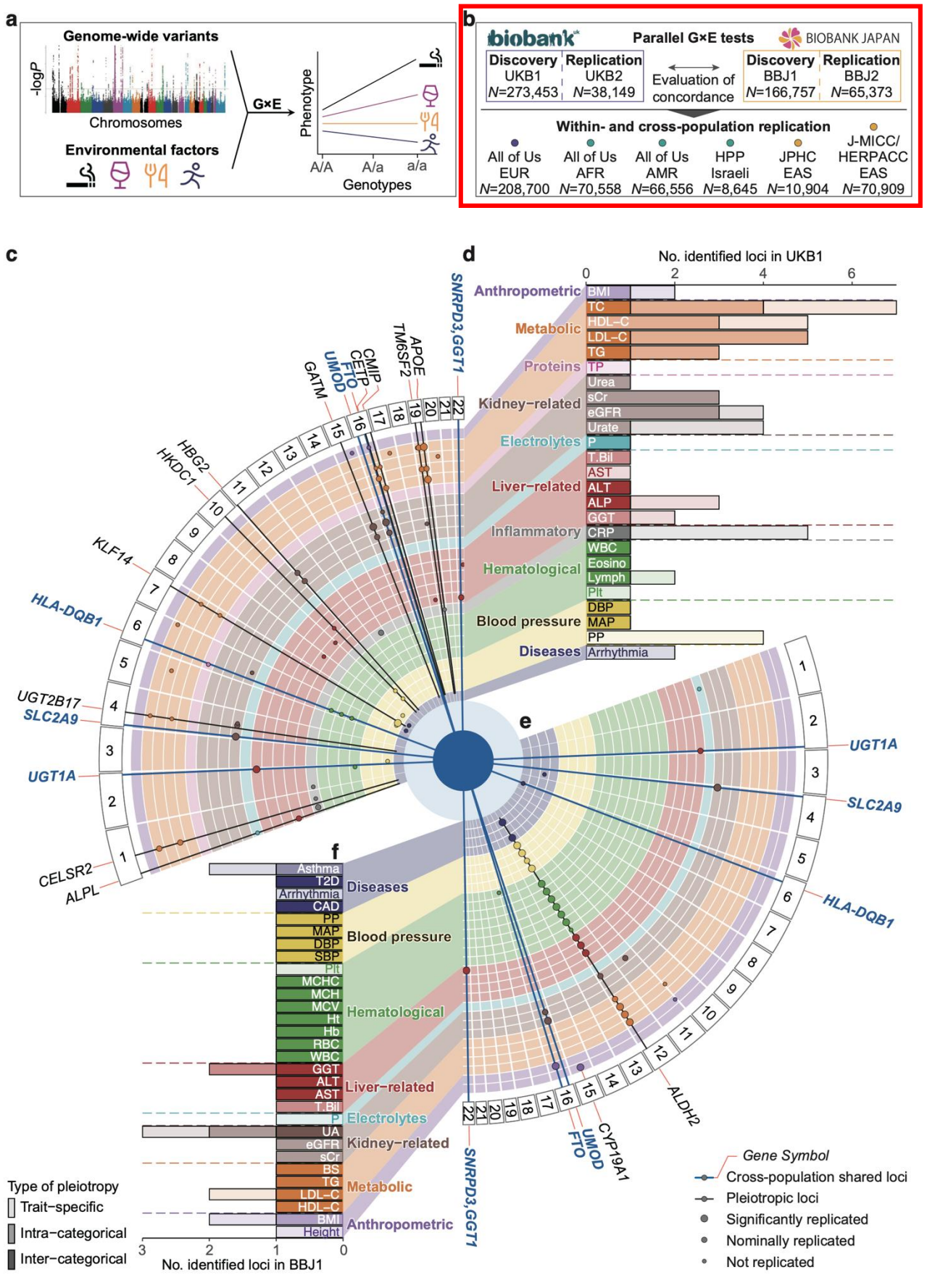
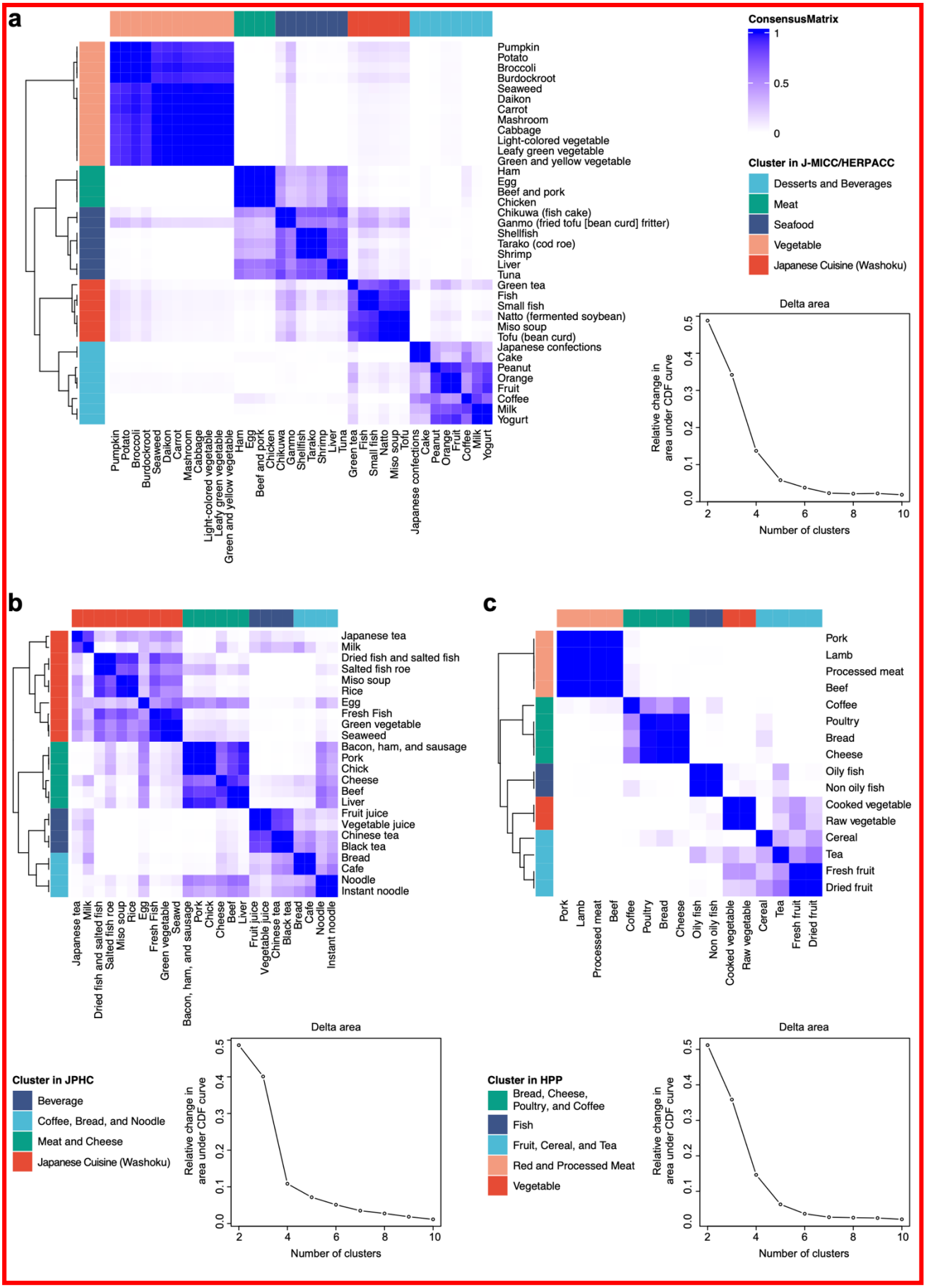

a
Genome-wide variants
-logP
Chromosomes
Environmental factors
G×E
Phenotype
A/A
A/a
a/a
Genotypes
b
biobank
Parallel G×E tests
BIOBANK JAPAN
Discovery
UKB1
N=273,453
Replication
UKB2
N=38,149
Evaluation of concordance
Discovery
BBJ1
N=166,757
Replication
BBJ2
N=65,373
Within- and cross-population replication
All of Us EUR N=208,700
All of Us AFR N=70,558
All of Us AMR N=66,556
HPP Israeli N=8,645
JPHC EAS N=10,904
J-MICC/ HERPACC EAS N=70,909
c
d
No. identified loci in UKB1
Anthropometric
Metabolic
Proteins
Kidney−related
Electrolytes
Liver−related
Inflammatory
Hematological
Blood pressure
Diseases
BMI
TC
HDL−C
LDL−C
TG
TP
Urea
sCr
eGFR
Urate
P
T.Bil
AST
ALT
ALP
GGT
CRP
WBC
Eosino
Lymph
Plt
DBP
MAP
PP
Arrhythmia
SNRPD3,GGT1
APOE
TM6SF2
CMIP
CETP
FTO
UMOD
GATM
HBG2
HKDC1
KLF14
HLA-DQB1
UGT2B17
SLC2A9
UGT1A
CELSR2
ALPL
e
UGT1A
SLC2A9
HLA-DQB1
ALDH2
CYP19A1
FTO
UMOD
SNRPD3,GGT1
f
Asthma
T2D
Arrhythmia
CAD
Diseases
PP
MAP
DBP
SBP
Blood pressure
Plt
MCHC
MCH
MCV
Ht
Hb
RBC
WBC
Hematological
GGT
ALT
AST
T.Bil
Liver−related
P
Electrolytes
UA
eGFR
sCr
Kidney−related
BS
TG
LDL−C
HDL−C
Metabolic
BMI
Height
Anthropometric
No. identified loci in BBJ1
Type of pleiotropy
Trait−specific
Intra−categorical
Inter−categorical
Gene Symbol
Cross-population shared loci
Pleiotropic loci
Significantly replicated
Nominally replicated
Not replicated

Figure 1, A cross-population atlas of G×E interactions

a, a schematic illustration of gene–environment (G×E) interactions. The icons were cited from Phosphor (https://phosphoricons.com/) under MIT license. **b**, an overview of the cohorts used in this study. EUR, AFR, AMR, and EAS denote the European, African, American, and East Asian populations, respectively. **c**, a circular plot of genome-wide significant G×E interactions in UK Biobank (UKB). The dot sizes represent the *P*-values of replication in UKB2 and BBJ2. Pleiotropic loci are shown with the lines and the nearest genes to the lead variants. **d**, the number of genome-wide significant loci for individual traits in UKB. The colors represent the trait categories, and the darkness of the bar colors represents the types of pleiotropy. **e** and **f**, the same as **c** and **d**, respectively, but for Biobank Japan (BBJ). The designs of the figures were inspired by and followed after the referenced studies.

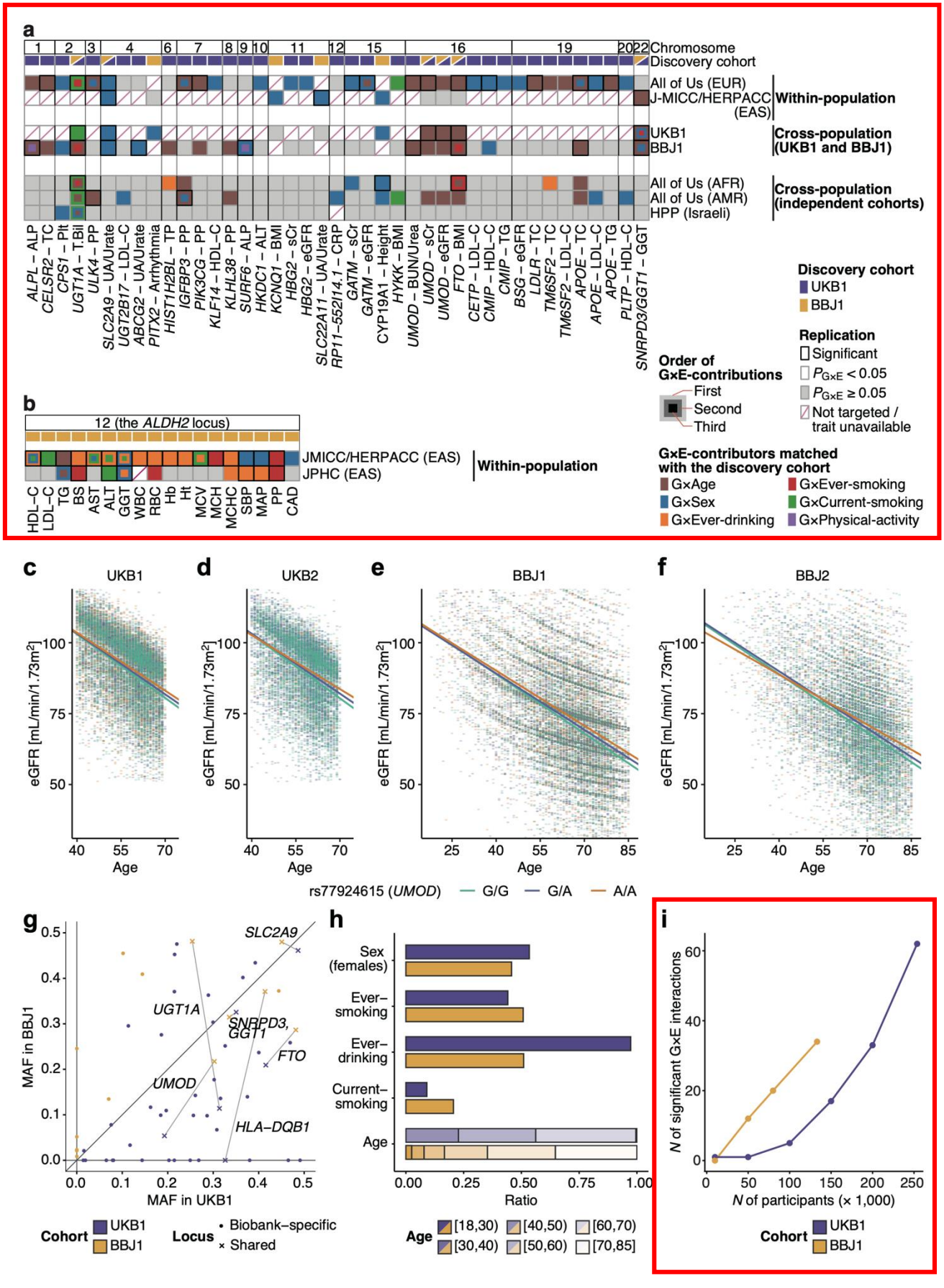
"We further tested replication in independent cohorts: the European population in All of Us (N_{max} = 208,700) and the East Asian population in two Japanese cohorts, J-MICC/HERPACC (N_{max} = 70,909) and JPHC (N_{max} = 10,904), for testing the G×E interactions detected in UKB1 and BBJ1, respectively (**Supplementary Table 10**). Despite the difference in the lifestyle questionnaire between cohorts, clustering of food frequency questionnaires detected dietary clusters consistent to those in the discovery cohorts, including the "Japanese cuisine (Washoku)" cluster observed across all Japanese cohorts and the "meat and cheese" cluster in UKB1 and JPHC, supporting the robustness of our clustering approach (**Ext. Data Fig. 4**). The Bonferroni replication rates was 27% in All of Us (17/62 trait–locus pairs) and 56% in J-MICC/HERPACC (20/36) ($P_{G\times E} < 5.1\times10^{-4}$; **Ext. Data Fig. 5a and b; Supplementary Table 11**). Notably, the pleiotropic G×E interactions at the *ALDH2* locus were replicated at 81% (17/21) in J-MICC/HERPACC, and six of them were again significantly replicated in JPHC, despite its relatively small sample size." (**Results**, P11–12 line 246–258).


a
ConsensusMatrix
1
0.5
0
Pumpkin
Potato
Broccoli
Burdockroot
Seaweed
Daikon
Carrot
Mashroom
Cabbage
Light-colored vegetable
Leafy green vegetable
Green and yellow vegetable
Ham
Egg
Beef and pork
Chicken
Chikuwa (fish cake)
Ganmo (fried tofu [bean curd] fritter)
Shellfish
Tarako (cod roe)
Shrimp
Liver
Tuna
Green tea
Fish
Small fish
Natto (fermented soybean)
Miso soup
Tofu (bean curd)
Japanese confections
Cake
Peanut
Orange
Fruit
Coffee
Milk
Yogurt
Cluster in J-MICC/HERPACC
Desserts and Beverages
Meat
Seafood
Vegetable
Japanese Cuisine (Washoku)
Delta area
Relative change in area under CDF curve
Number of clusters
b
Japanese tea
Milk
Dried fish and salted fish
Salted fish roe
Miso soup
Rice
Egg
Fresh Fish
Green vegetable
Seawed
Bacon, ham, and sausage
Pork
Chick
Cheese
Beef
Liver
Fruit juice
Vegetable juice
Chinese tea
Black tea
Bread
Cafe
Noodle
Instant noodle
Cluster in JPHC
Beverage
Coffee, Bread, and Noodle
Meat and Cheese
Japanese Cuisine (Washoku)
Delta area
Relative change in area under CDF curve
Number of clusters
c
Pork
Lamb
Processed meat
Beef
Coffee
Poultry
Bread
Cheese
Oily fish
Non oily fish
Cooked vegetable
Raw vegetable
Cereal
Tea
Fresh fruit
Dried fruit
Cluster in HPP
Bread, Cheese, Poultry, and Coffee
Fish
Fruit, Cereal, and Tea
Red and Processed Meat
Vegetable
Delta area
Relative change in area under CDF curve
Number of clusters

Extended Data Figure 4. Consensus clustering of dietary questionnaires in independent replication cohorts.

Consensus clustering of food frequency questionnaires in J-MICC/HERPACC (a), JPHC (b), and HPP (c). Delta area plots show the area under the cumulative distribution function (CDF) curve when changing the number of clusters from two to ten.

"We further evaluated replication in the African and American populations in All of Us (N_{max} = 70,558 and 66,556, respectively) and the Israeli population in the Human Phenotype Project (HPP; N_{max} = 8,645). In the African population, three G×E interactions were significantly replicated, two of which overlapped with the shared G×E interactions between UKB1 and BBJ1 (**Ext. Data Fig. 5a and Supplementary Table 13**; $P_{G\times E} < 0.05/77/3 = 2.3\times10^{-4}$, corrected for the number of trait-locus-biobank triplets outside of the *ALDH2* locus and the number of populations). In the American population, two G×E interactions for pulse pressure (PP) were significantly replicated, both originally from UKB1 and primarily driven by G×Age. G×E interactions were not significantly replicated in the Israeli population, possibly due to its relatively small sample size. Nevertheless, we observed the top signal for one of the shared G×E interactions between UKB1 and BBJ1 (*UGT1A* gene family for T.Bil, commonly driven by G×Current-smoking and G×Age)." (**Results**, P13 line 296–307).



Extended Data Figure 5. Replication within and across populations.

a–b, Replication statuses of G×E loci other than the *ALDH2* locus (**a**) and the *ALDH2* locus (**b**). Only interactions with at least one nominally significant replication ($P_{G\times E} < 0.05$) are shown. EUR, the European population; EAS, the East Asian population; AFR, the African population; AMR, the American population. **c–f**, eGFR distributions across rs77924615 genotypes (the *UMOD* locus) and age in UKB1 (**c**), UKB2 (**d**), BBJ1 (**e**), and BBJ2 (**f**), as a representative example of cross-population shared G×E interaction. Dots, 5,000 randomly sampled individuals per genotype; lines, regression lines using all participants. **g and h**, minor allele frequencies (MAF) of the lead variants (**g**) and environmental distributions (**h**) in UKB1 and BBJ1. In **g**, cross-population shared loci are marked with crosses and labeled by nearest genes. Solid gray lines represent $x=0$, $y=0$, and $y=x$. **i**, number of G×E loci detected with $P_{G\times E}$ less than 5.0×10^{-8} in randomly subsampled participants in UKB1 and BBJ1. Subsample sizes: 10, 50, 100, 150, and 200 thousand in UKB1; 10, 50, and 80 thousand in BBJ1. Full-cohort results also shown (N_{mean} = 253,773 in UKB1 and 133,117 in BBJ1).

We described the overview of the replication cohorts in **Supplementary Methods**.

"**Japan Multi-Institutional Collaborative Cohort (J-MICC) Study and Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC)**

J-MICC is a multi-institutional collaboration of prospective and population-based cohorts comprising approximately 100,000 participants in Japan, launched in 2005 with the primary purpose of detecting G×E interactions of lifestyle-related diseases. Individuals aged 35 to 69 years were enrolled from 2005 to 2010 from respondents to study announcements in specified areas, inhabitants attending health checkup examinations by local governments, visitors at health checkup centers, and patients at a cancer hospital. The second assessment was conducted five years after their enrollment. HERPACC is a hospital-based cohort including participants aged 18 or older recruited from 2001 to 2013 from first-visit outpatients at Aichi Cancer Center Hospital. Aichi Cancer Center Hospital is one of the constitutive cohorts of J-MICC, and the same questionnaire was used for both J-MICC and HERPACC. After removing duplicated participants between J-MICC and HERPACC, we analyzed the participants of J-MICC and HERPACC together, which we refer to as J-MICC/HERPACC. For HERPACC, we analyzed the participants with no cancer or history of neoplasm to match the cohort to the population-based characteristics of J-MICC." (**Supplementary Methods**).

"**Japan Public Health Center-based Prospective (JPHC) Study**

The JPHC study is a population-based cohort comprising participants enrolled from residents aged 40-69 years within 11 public health center (PHC) areas nationwide between 1990 and 1994. Before conducting genetic research in the JPHC study, we obtained approval from the institutional review board of the National Cancer Centre (Approval No.: 2011-044), Tokyo, Japan, and provided eligible participants with the option to decline participation. For the present study, 13,024 participants were randomly selected as a subcohort from the JPHC study participants in 9 PHC areas, who had received a health check-up and answered a questionnaire during the baseline survey." (**Supplementary Methods**).

"**All of Us**

All of Us is a nation-wide cohort in the United States, aiming to recruit persons in demographic categories that have been and continue to be underrepresented in biomedical research.

National enrolment began in 2018, and adults 18 years and older who have the capacity to consent and reside in the United States or the United States territory at present are eligible. We analyzed the version 8 dataset of the All of Us." (**Supplementary Methods**).

"**The Human Phenotype Project (HPP)**

HPP is a prospective cohort in Israel, and the participants between the ages of 40 and 70 years were recruited from 2018 to 2021 from the self-assignment of volunteers registered to the trial website. The cohort is largely composed of Ashkenazi Jewish people. Individuals without recent antibiotic use and gastrointestinal morbidity were eligible if they had not yet been diagnosed with the clinical outcomes of interest." (**Supplementary Methods**).

"**Quality control of phenotypes and environments in the replication cohorts**

Clinical traits

We employed the same definition and quality control for clinical traits as those used in UKB1 and BBJ1. In All of Us, biomarker phenotypes were obtained from the electronic medical records, and we used the most recent records for each biomarker. When both medical records from the emergency department and those from other departments were available, we prioritized the latter. We observed frequent mislabeling of the measurement units in this cohort. For example, albumin measurements with the units of mg/mL and mg/L were of the same order of magnitude as those reported in g/dL. Therefore, we corrected the units by examining the median values for individual unit labels. The biomarker phenotypes for the other cohorts were obtained from the baseline assessment data for the main analyses, while we used the second assessment data of J-MICC for the temporal order analyses. We applied the same quality control criteria as used for BBJ and UKB. For disease statuses, we defined participants as cases if they had relevant SNOMED CT codes or self-reported disease histories in All of Us. We used self-reported disease histories for the other cohorts. For JPHC and J-MICC/HERPACC, self-reported histories were recorded for diabetes mellitus instead of T2D, and we used diabetes mellitus as a proxy of T2D for replication purposes.

Environmental factors

The environmental factors from questionnaires, along with the conversion of categorical responses into a continuous scale, were summarized in **Supplementary Table 4**. As only one environmental factor was related to physical activity in JPHC, clustering it with dietary environmental factors would lead to a cluster consisting of both physical activity and dietary factors, opposite to our observation that the "physical activities" cluster was consistently detected as a separate cluster from other dietary clusters in UKB1 and BBJ1. Therefore, we pre-defined the "physical activities" cluster from the environmental factors related to physical activities in the replication cohorts and performed clustering only using dietary factors. In JPHC, two sets of questionnaires with slight differences in questionnaire items were used, and we used the items available in both sets. Questionnaires related to physical activities and dietary consumption were unavailable in All of Us, and we did not perform clustering in this cohort. In All of Us, smoking status was separately asked for cigar smoking, electronic

smoking, hookah smoking, and smokeless tobacco smoking. We defined ever- and current-smoking statuses based on those of all types of tobacco." (**Supplementary Methods**).

2) Clustering dietary survey responses to define exposure groups/clusters. How are those meaningful to interpretation? For example, how can you use the results to provide a recommendation for reducing or increasing consumption of a dietary product in the context of genetic background if you don't know what amount/level of consumption matters because the measure of exposure you detected was a group based assignment of individuals with similar questionnaire/survey response profiles (principal components) instead of the value of the exposure itself? And could the environment clusters/groups be biased/clustering by some other factor related to how they answered the questionnaire?

Response:

Thank you for your insightful comment. We agree with you that interpretation and clinical applicability are important considerations when using clustering-based representations of environmental factors.

We employed a clustering approach to group dietary and physical activity questionnaires, following the strategy in Westerman KE *et al.* ***Hum Mol Genet*** (2021), with the aim of enhancing the detection power of G×E interactions and reducing multicollinearity among highly correlated questionnaire items. The cluster scores were standardized to a mean of zero and a standard deviation of one and used as quantitative variables in the G×E interaction models.

This approach was particularly helpful in identifying the environmental factor that contributed to the G×E interaction at the *ABCG2* locus, where meat consumption consistently drove the G×E interaction across meat types. In this case, the cluster-level variable provided a better interpretation of the G×E interaction than analyzing individual food items separately. On the other hand, we reported the G×Natto interaction at the *PITX2* locus by inspecting the raw questionnaire items. These examples demonstrate that **cluster-level and item-level analyses can complement each other in facilitating the interpretation of G×E interactions**.

Regarding your concern about potential bias in clustering, we refer to Mignogna G *et al*, ***Nat Hum Behav*** (2023), who reported the association between item nonresponse behavior and socioeconomic status. They reported that the item nonresponse behavior was highly similar among survey questions from similar phenotype domains; for example, they reported that the average correlation (r^2) of the "prefer not answer" answers among questions within the food intake domains was 0.85 (IQR = 0.08). Based on this, we believe the risk of spurious clustering within dietary domains was limited. In addition, we observed similar dietary clusters in independent cohorts, including the "Japanese cuisine (Washoku)" cluster observed across all three Japanese cohorts, and the "meat and cheese" cluster observed in both UKB and JPHC. These observations support the robustness and reproducibility of our clustering approach.

Modification:

We described the clustering results of the questionnaires in the independent replication cohorts in **Results** and **Ext. Data Fig. 4**, as described in Comment #1-2. We also described the standardization of the cluster scores in **Methods**.

"The cluster scores were standardized to a mean of zero and a standard deviation of one." (**Methods**, P41 line 1036–1037).

3) The results report some loci show GxE (e.g. ALDH2) associations with multiple environments and argue this provides biologic mechanism insights. I agree biologic insights can be gained but is it also possible some of the same GxE across multiple exposures could be due to exposure correlation?

Response:

Thank you for your thoughtful comment. We agree with you that some environmental factors were still correlated after clustering similar questionnaires. To account for these correlations, in both initial and revised manuscripts, we employed an iterative approach to determine the environments that contributed to the G×E interactions. Briefly, we added G×E interaction terms one by one to the regression model in the order of their importance and retained them only if the model fit was improved with a *P*-value less than 0.05 (likelihood ratio test). In principle, this approach selected more parsimonious sets of environments than the Akaike information criterion, and it ensures that each selected environment explains additional variance in the G×E effect conditional on previously selected environments. Therefore, **if multiple environments were selected, it indicates that they provided independent explanatory values, despite any correlation among them**.

We are also aware that identifying the top environment contributing to G×E interaction is important to obtain insights into the underlying mechanisms. Therefore, we reported the G×E-contributing environments in descending order of their contributions in the Supplementary Tables. This approach allowed us to highlight the G×Natto interaction at the *PITX2* locus for arrhythmia. Other environments also contributed to this locus, including sex, ever-drinking, and ever-smoking. Nevertheless, testing the G×Natto interaction alone surpassed the genome-wide significance threshold ($P_{G\times E} = 2.1\times10^{-10}$), supporting the primary contribution of G×Natto interaction at this locus. A detailed discussion on this locus can be found in our response to Comment #5. Variable selection based on the Akaike information criteria (AIC) does not allow for identifying the top-contributing environment, providing a further rationale for our approach.

Modification:

We revised the order of sentences in the relevant **Methods** section to further enhance readability and to describe the rationale of our approach more clearly.

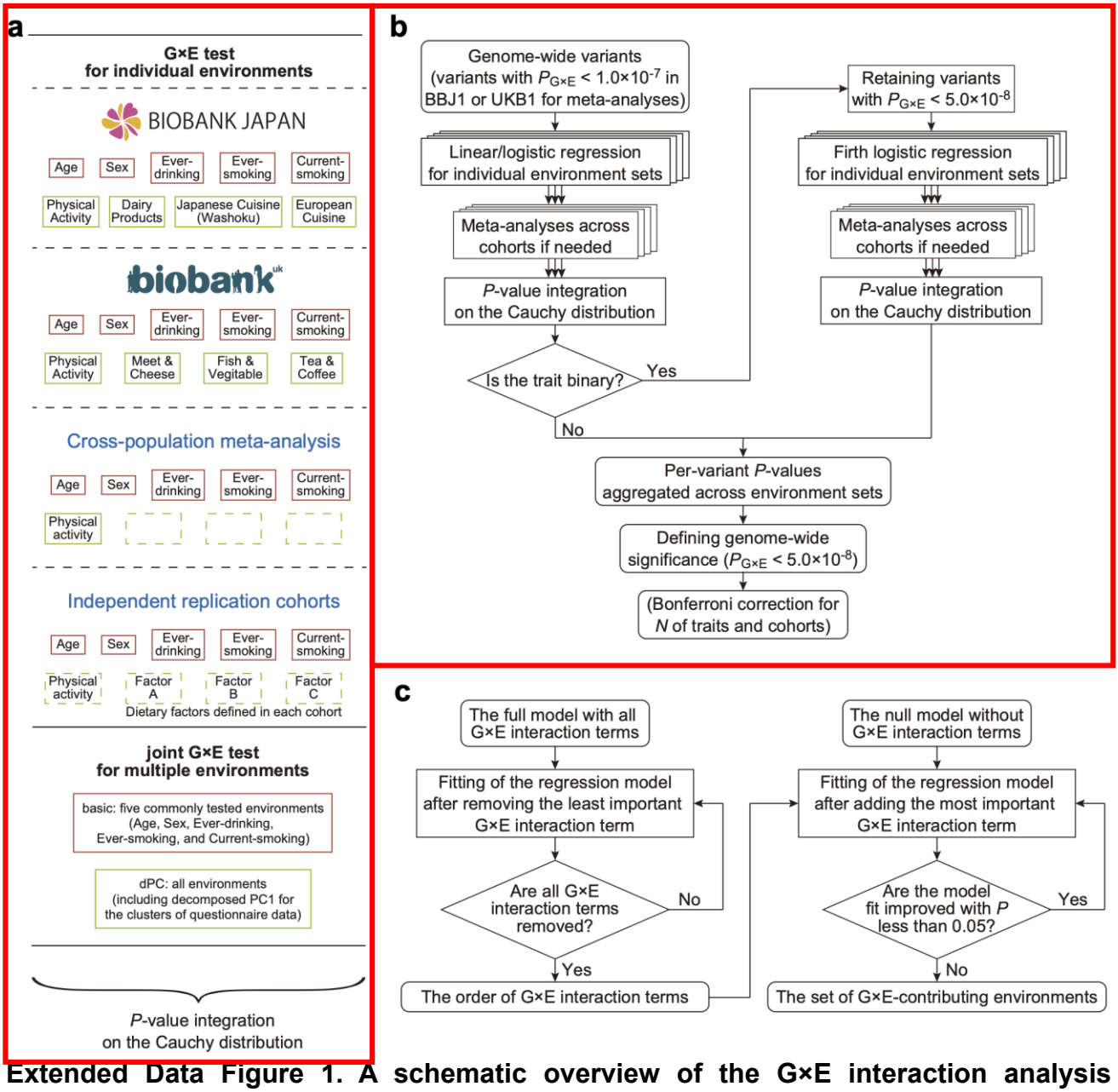
"We tested G×E interactions for different sets of environmental factors, as shown in **Ext. Data Fig. 1**, and then aggregated *P*-values across environmental sets on the Cauchy distribution[31] to obtain per-variant *P*-values. When *P*-values were 0, which meant *P*-values were less than 1.0×10^{-300}, we estimated the precise *P*-values using the multiple-precision floating-point arithmetic with the maximal precision of 10,000 bits, implemented in the Rmpfr R package. We defined a G×E interaction as significant at the variant level if the aggregated *P*-value was

below the genome-wide threshold of 5.0×10^{-8}. We also reported the study-wide Bonferroni threshold for full transparency. We did not require that any individual environment's interaction term be significant on its own, considering that G×E interactions can be driven by multiple environments. Nevertheless, we note that at least one raw *P*-value was smaller than the Cauchy-combined *P*-value in principle, as the Cauchy combination method is not a meta-analysis but rather returns a *P*-value within the range of raw *P*-values.

After evaluating the significance of G×E interactions at the variant level, we determined the order of importance of G×E interaction terms by backward elimination from the regression model with all G×E interaction terms. Specifically, we repeatedly removed the G×E interaction term with the least improvement in the likelihood of the linear regression model or the largest Wald *P*-value of the Firth logistic regression model. After removing all G×E interaction terms, we brought back the G×E interaction terms one by one in the order of their importance if the model fit was improved with a *P*-value less than 0.05 (likelihood ratio test). We considered that the environmental factors contributed to G×E interactions if the corresponding interaction terms were included in the final model. In principle, this approach selects more parsimonious sets of environments than the Akaike information criterion and can determine the top environment contributing to individual G×E interactions.

" (**Methods**, P43–44 line 1090–1112).

We also revised **Ext. Data Fig. 1** (**Supplementary Figure 1** in the initial manuscript) to represent our approach more clearly.



Extended Data Figure 1. A schematic overview of the G×E interaction analysis workflow.

a, The environment sets used to test G×E interactions. *P*-values were calculated for individual environmental sets and integrated on the Cauchy distribution to obtain final *P*-values. **b**, The workflow of G×E interaction tests. **c**, the workflow of defining the environments which contributed to G×E interactions and their orders.

4) Are these clinically meaningful differences in the effects of E by G? They reach statistical significance but what is the absolute or relative risk difference by G? For example, in Supplementary Figure 4, there seems to be a very small difference in eGFR change by genotype even in the oldest age, around 70 years.

Response:

Thank you for your important comment. We acknowledge that the absolute effect sizes of the G×E interactions for clinical traits were generally modest. **This is consistent with findings from GWAS, where common variants often have small individual effect sizes, yet their aggregated effects, represented by PGS, can have large effect sizes**.

In our manuscript, we reported that the G×E heritability was significantly positive for 18 biobank-trait pairs at the genome-wide level, with their mean G×E heritability of 3.26%. As discussed in the response to Comment #1-1, the heritability determines the upper bound for the prediction accuracy that an ideal PGS could theoretically achieve (O'Connor LJ ***Nat Genet*** (2021)). Therefore, our results suggest that incorporating G×E interactions may have the potential to improve the prediction accuracy of PGS at a clinically meaningful level in theory, although methodological advancements are warranted to fully realize this potential. As we described in response to #1-1, the G×E-PGS constructed from G×Sex interactions indeed improved the prediction accuracy of BMI.

Beyond prediction, it has been reported for GWAS that the support by genetic associations doubles the success rate of drug development, irrespective of their effect sizes (Minikel EV *et al.* ***Nature*** (2024)). G×E interactions may also offer insights into therapeutic development. We indeed reported the sexually dimorphic association at the *CETP* locus, which may be informative in explaining the failure to develop drugs targeting this gene.

In addition to the *CETP* locus, we observed multiple loci that showed sexually dimorphic associations with opposite directions of effect sizes between sexes, the extreme form of G×E interactions, for the metabolome. Motivated by these G×E interactions, **we additionally conducted genome-wide G×E interaction tests for the 325 metabolites. We note that we expanded the sample sizes of BBJ1 for metabolome from 43,371 to 89,040**, substantially enhancing the power to detect G×E interactions. We identified 30 and 11 genome-wide significant loci in UKB1 and BBJ1, respectively, yielding 736 and 228 metabolite–locus pairs. Among these, 16 loci in UKB1 and 4 in BBJ1 passed the Bonferroni-corrected threshold ($P_{G\times E} < 7.7\times10^{-11}$, correcting for 325 metabolites across two cohorts). Notably, G×Sex interactions at the *ALDH1A2* and *ZNF259* loci were detected in both cohorts. In addition, **several loci exhibited sex-dimorphic G×Sex interactions, with opposite directions of association between males and females**, including the *LPL*, *FADS2/FADS1*, *APOB*, and *GLDC* loci, further underscoring the importance of sex-specific genetic architecture in metabolome regulation. Although not reaching the stringent Bonferroni threshold, we also reported notable G×E interactions, included a current-smoker-specific association at the *LRRC4C* locus for phospholipids to total lipids ratio in large VLDL (L_VLDL_PL_pct, $P_{G\times E} = 1.1\times10^{-8}$), a never-drinker-specific association at the *KLHL32* locus for phospholipids in small VLDL (S_VLDL_PL, $P_{G\times E} = 4.8\times10^{-9}$), and a G×E interaction without a marginal effect for acetone ($P_{G\times E} = 2.2\times10^{-8}$) at the *SLC7A2* locus, driven by tea and coffee consumption, ever smoking, and meat and cheese consumption. **Collectively, our systematic evaluation of omics G×E interactions suggested that studying endophenotypes, such as the metabolome, may detect G×E interactions with larger effect sizes than clinical traits**.

Modification:

We added the genome-wide G×E interaction tests for the metabolome as follows.

"Motivated by these G×E metabolite QTLs, we expanded the metabolome-wide G×E analysis to genome-wide variants, identifying 30 and 11 genome-wide significant loci in UKB1 and BBJ1, respectively, yielding 736 and 228 metabolite–locus pairs. Among these, 16 loci in

UKB1 and 4 in BBJ1 passed the Bonferroni threshold ($P_{G\times E} < 7.7\times10^{-11}$, correcting for 325 metabolites across two cohorts). Notably, G×Sex interactions at the *ALDH1A2* and *ZNF259* loci were observed in both cohorts (**Supplementary Table 31**). Several loci exhibited sex-dimorphic G×Sex interactions with opposite effect directions in males and females. These included the *LPL*, *FADS2/FADS1*, *APOB*, and *GLDC* loci (**Ext. Data Fig. 10**; $P_{G\times E} = 8.7\times10^{-24}$, 7.7×10^{-27}, 6.6×10^{-17}, and 4.9×10^{-28}, respectively), again underscoring the sex-specific genetic architecture of metabolome regulation. Additional notable G×E signals, though below Bonferroni significance, included a current-smoker-specific association at the *LRRC4C* locus for phospholipids to total lipids ratio in large VLDL (L_VLDL_PL_pct; $P_{G\times E} = 1.1\times10^{-8}$), and a never-drinker-specific association at the *KLHL32* locus for phospholipids in small VLDL (S_VLDL_PL; $P_{G\times E} = 4.8\times10^{-9}$). At the *SLC7A2* locus, we observed marginal effects for alanine and glutamine and a G×E-only effect for acetone ($P_{G\times E} = 2.2\times10^{-8}$), driven by tea and coffee consumption, ever smoking, and meat and cheese consumption (**Ext. Data Fig. 10**). Further validation and functional studies are warranted to elucidate their underlying mechanisms." (**Results**, P25–26 line 615–632).

"For the genome-wide G×E interaction tests for metabolome, we targeted (i) genotyped variants in each cohort, (ii) HapMap3 variants, and (iii) the variants targeted by the meta-analysis across BBJ and UKB, to reduce the computational cost. After defining the metabolome G×E loci based on their distance as above, we merged the metabolome G×E loci if they overlapped with the same locus defined for the clinical traits." (**Methods**, P44–45 line 1124–1129)

We also discussed that studying omics data may detect G×E interactions with large effect sizes.

"We further conducted omics-level G×E QTL analyses to pinpoint the metabolites and proteins that mediate clinical G×E interactions. Multiple sex-dimorphic and environment-specific genetic effects on the metabolome suggest that omics data may reveal G×E interactions with large effect sizes." (**Discussion**, P27–28 line 660–665).

5) The results section on "A disease-induced dietary change caused a GxE interaction" is particularly important, I think, to highlighting the causality and directionality of GxE is not always upstream of the outcome. This may be something that the authors want to state in the discussion/highlight as an important consideration with GxE result interpretation, generally. I think many in the field will assume it is the gene-environment interaction causing the outcome and don't always consider that gene-outcome/treatment can cause people to change their "environment" and lead to a "GxE".

Response:

Thank you for acknowledging the importance of this section. Your comment was truly helpful and motivated us to conduct additional analyses to further validate the reverse causality of this G×E interaction. First, **we newly observed that natto consumption significantly decreased after the initiation of warfarin treatment, based on within-individual comparisons in longitudinal data** (**Fig. 2i**). The causal direction from the disease to natto

consumption was further supported by Mendelian randomization between warfarin treatment and natto consumption ($P = 6.6\times10^{-3}$). A Cox regression analysis using incidental cases in BBJ1 was not significant at this locus ($P_{G\times E} = 0.077$), which was expected given that the Cox regression model tested the temporal direction from environments to diseases, opposite to the temporal order observed above. We also note that testing the G×Natto interaction alone surpassed the genome-wide significant threshold ($P_{G\times E} = 2.1\times10^{-10}$), supporting the primary contribution of G×Natto interaction at this locus.

Furthermore, we found that the marginal effect size of the lead variant was heterogeneous between the subgroups of arrhythmia, making the overall effect size for arrhythmia dependent on the rate of atrial fibrillation/flutter patients among the arrhythmia patients. This led to a difference in the effect size for arrhythmia across strata of natto consumption frequency, which explained the link between the reverse causality and the G×Natto interaction. In line with this notion, the G×Natto interaction was not significant when evaluated separately within the atrial fibrillation/flutter and the other arrhythmia subgroups ($P_{G\times E} = 0.84$ and 2.6×10^{-4}, respectively).

Together, these results support the reverse causal relationship from the disease to the environment, consistent with clinical practice in Japan, where physicians commonly advise warfarin-taking patients to avoid eating natto.

Modification:

We added the additional evidence to support the reverse causal direction in **Results**.

"Testing G×Natto alone surpassed genome-wide significance ($P_{G\times E} = 2.1\times10^{-10}$), supporting its primary role at this locus." (**Results**, P15 line 350–351).

"Indeed, natto intake declined markedly after warfarin initiation in the same individuals (**Figure 2i**). Mendelian randomization supported the negative causal effect of warfarin on natto consumption ($P = 6.6\times10^{-3}$, effect size = -0.029 [95% confidence interval (CI), -0.051 to -8.2×10^{-3}]).

When stratifying arrhythmia patients by clinical subgroups, the marginal effect size of rs72900155 was substantially larger for atrial fibrillation/flutter than for other subgroups (0.49 [95% CI, 0.46–0.52] versus 0.13 [0.10–0.15]). Due to this effect size heterogeneity, the overall effect size for arrhythmia would depend on the proportion of atrial fibrillation/flutter patients and therefore differ across natto intake strata, explaining the link between the reverse causality and the G×Natto interaction. Consistently, the G×Natto interaction was not significant within either subgroup when evaluated separately ($P_{G\times E} = 0.84$ for atrial fibrillation/flutter; 2.6×10^{-4} for other subgroups)." (**Results**, P16 line 360–371).

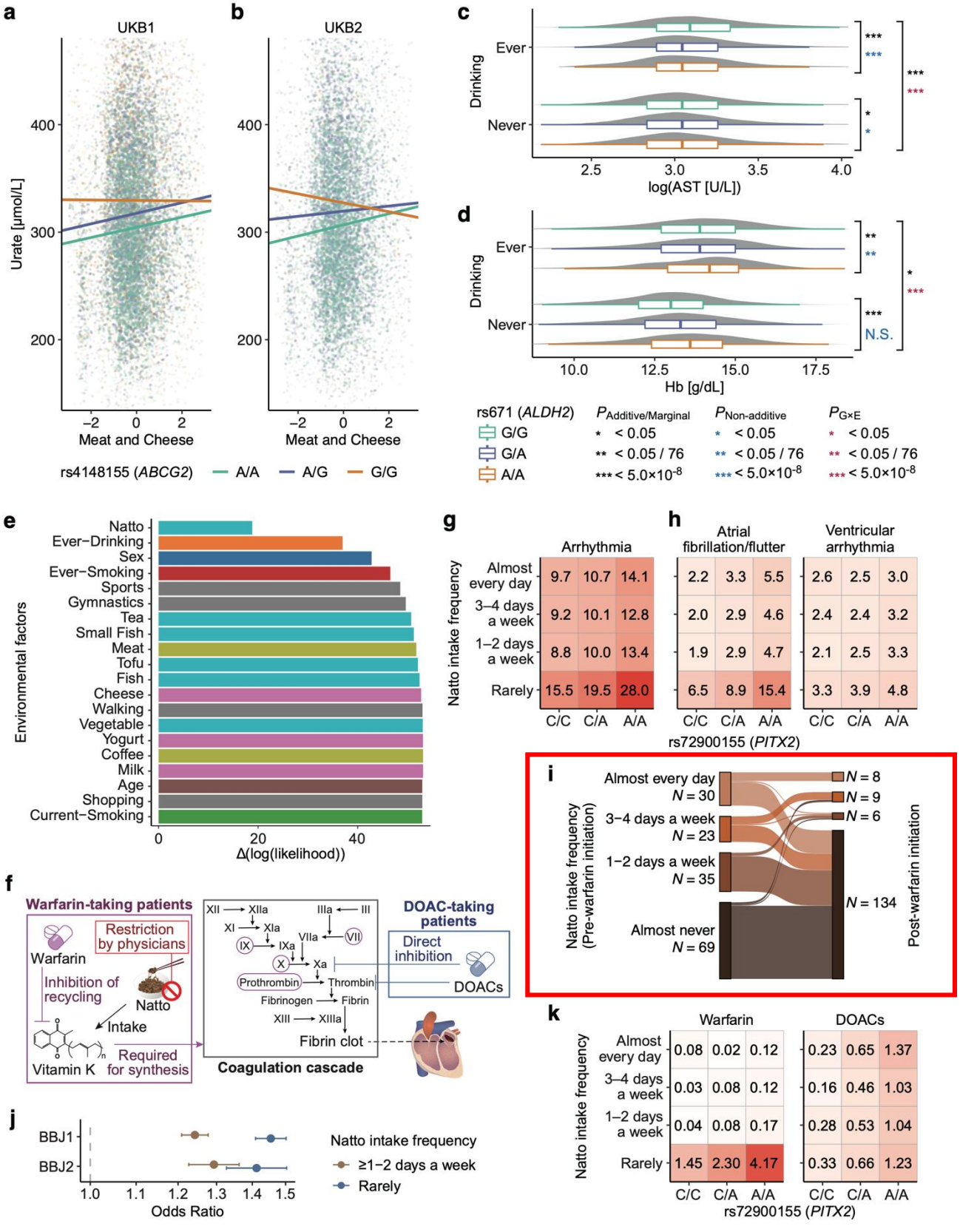


Figure 2, Representative loci with G×E interactions.

a and **b**, Urate levels across rs4148155 genotypes (the *ABCG2* locus) and "meat and cheese" consumption in UKB1 (**a**) and UKB2 (**b**). Dots show 5,000 randomly sampled individual measurements per genotype; lines represent the linear regression lines using all participants. **c** and **d**, AST (**c**) and hemoglobin (Hb, **d**) levels by rs671 genotypes (the *ALDH2* locus) and drinking status in BBJ1. Boxplots show median, quartiles, and 1.5× interquartile range whiskers. **e**, Log-likelihood improvement for G×E models of rs72900155 (the *PITX2* locus) for arrhythmia. Environmental factors were added stepwise to the null model (**Methods**). **f**, schematic of the mechanisms of action for warfarin, natto (fermented soybean), and direct oral anticoagulants (DOAC). Heart icon from TogoTV (© 2016 DBCLS TogoTV, CC-BY-4.0). **g**, heatmap of the arrhythmia prevalence in BBJ1. **h**, same as **g**, stratified by arrhythmia subgroups. **i**, natto intake before and after warfarin initiation in BBJ atrial fibrillation/flutter patients with ≥1 PT-INR record per year, considering that regular monitoring of PT-INR is required during warfarin therapy. **j**, odds ratios of rs72900155 (the *PITX2* locus) for arrhythmia stratified the natto intake frequency. Error bars represent 95% confidence intervals (CI; Firth logistic regression). **k**, heatmap of the atrial fibrillation/flutter prevalence in BBJ2, stratified by warfarin and DOAC medication.

We also described the Cox regression analysis using incidental cases in **Supplementary Note 8**.

"On the other hand, the G×E interaction for arrhythmia at the *PITX2* locus was not significant ($P_{G\times E}$ = 0.077). In the section "A disease-induced dietary change caused a G×E interaction," we discussed that this locus was driven by a reverse causal relationship from a disease to an environment (specifically, from warfarin medication for atrial fibrillation/flutter to the avoidance of natto consumption). Considering that the Cox regression model tested the temporal direction from environments to diseases, the failure of its verification in the Cox regression was expected." (**Supplementary Note 8**).

"All statistical tests were two-sided unless otherwise noted. As phenotypes, we targeted quantitative traits (clinical biomarkers, metabolites, and protein expression measurements), dichotomous traits (disease statuses), and time-to-event traits (disease onset for incidental cases and overall survival). For quantitative and dichotomous traits, we tested G×E interactions using GEM, a fast and scalable implementation of linear and logistic regressions with G×E interaction terms[23] (**Ext. Data Fig. 1**). We estimated the model-robust "sandwich" standard errors to suppress the inflation of statistics[23]. As the case–control imbalance can cause the inflation of test statistics in the logistic regression, we re-estimated the effect sizes and *P*-values using Firth logistic regression for the variants with *P*-values less than 5.0×10^{-8} for dichotomous traits. For the time-to-event traits, we employed the Cox proportional hazard model implemented in the R package "survival". Individuals who had already developed the target disease at baseline or developed it within six months from baseline were excluded from the incidental case analyses." (**Methods**, P43 line 1077–1089).

6) In the introduction section, this sentence isn't clear to me: "We further projected environment-stratified genetic effects to scRNA-seq data to uncover the dynamic shift of responsible cell types at single-cell resolution". I don't really understand what question you sought to ask or how you answered it methodologically from this sentence.

Response:

Thank you for your comment. We revised the sentence to make it clearer.

Modification:

According to your kind suggestion, we revised the sentence as follows:

"We further inferred the cell types responsible for the genetic architecture of pulse pressure at younger and older ages, uncovering dynamic age-dependent shifts in cell type involvement at single-cell resolution." (**Introduction**, P6 line 143–145).

7) In Supplementary Figure 3 – what is the reference population used for the r2 correlation values shown? European? What does the correlation structure look like in other global populations?

Response:

Thank you for your comment. We used the European in-sample linkage disequilibrium information from UKB1. We agree with you that the structure of linkage disequilibrium in the loci that appeared in the figure may differ across populations. However, these loci were not significantly replicated in non-European replication cohorts, limiting our ability to reliably assess the separation of these signals in non-European populations.

Modification:

We clarified that we used the European in-sample linkage disequilibrium information from UKB1 in the legend of the figure.

"**Extended Data Figure 3. LocusZoom plots of inter-categorical pleiotropic loci in UKB1.**

a-i, LocusZoom plots of the significant G×E interactions at the *ALPL* (**a and b**), *HKDC1* (**c–e**), and *APOE* (**f–i**) loci. **j–l**, hg19 coordinates of protein-coding genes in individual loci. Recombination rates were calculated using the European in-sample linkage disequilibrium information from UKB1." (The legend of **Ext. Data Fig. 3** [**Supplementary Figure 3** in the initial manuscript]).

8) Figure 3a-b reports GxE heritability metrics. It would be helpful to also see (for reference) what the heritability estimate for G and E alone would look like in comparison to these GxE estimates. I struggled a bit to understand the point here and how it would relate to my expectations of more "traditional" heritability estimates for outcomes assessing marginal G or E effects.

Response:

Thank you for your helpful comment. **We have added Figure 3c to more clearly illustrate the G×E heritability in comparison with the "traditional" heritability of marginal effects**.

Consistent with "traditional" heritability models, the variance explained by environmental factors, including unmeasured ones, is represented as (1 - h^2). For the variance explained by measured environmental factors, the corresponding concept in the G×E multivariate reaction norm model is the heterogeneous residual effects across environmental values (i.e., $\sigma^2_{R\times E}$), as described by Shin J and Lee SH ***Genome Biol*** (2021). Their theoretical framework indicated that $\sigma^2_{R\times E}$ could be estimated by subtracting $h^2_{G\times E}$ from the LDSC intercept. However, since the LDSC intercept generally reflects inflation of test statistics, we refrained from estimating $\sigma^2_{R\times E}$ to avoid reporting potentially biased values. Estimating the variance explained by measured environmental factors within the LDSC framework is challenging, and further methodological advancements are warranted.

Modification:

According to your kind suggestion, we added **Figure 3c**.

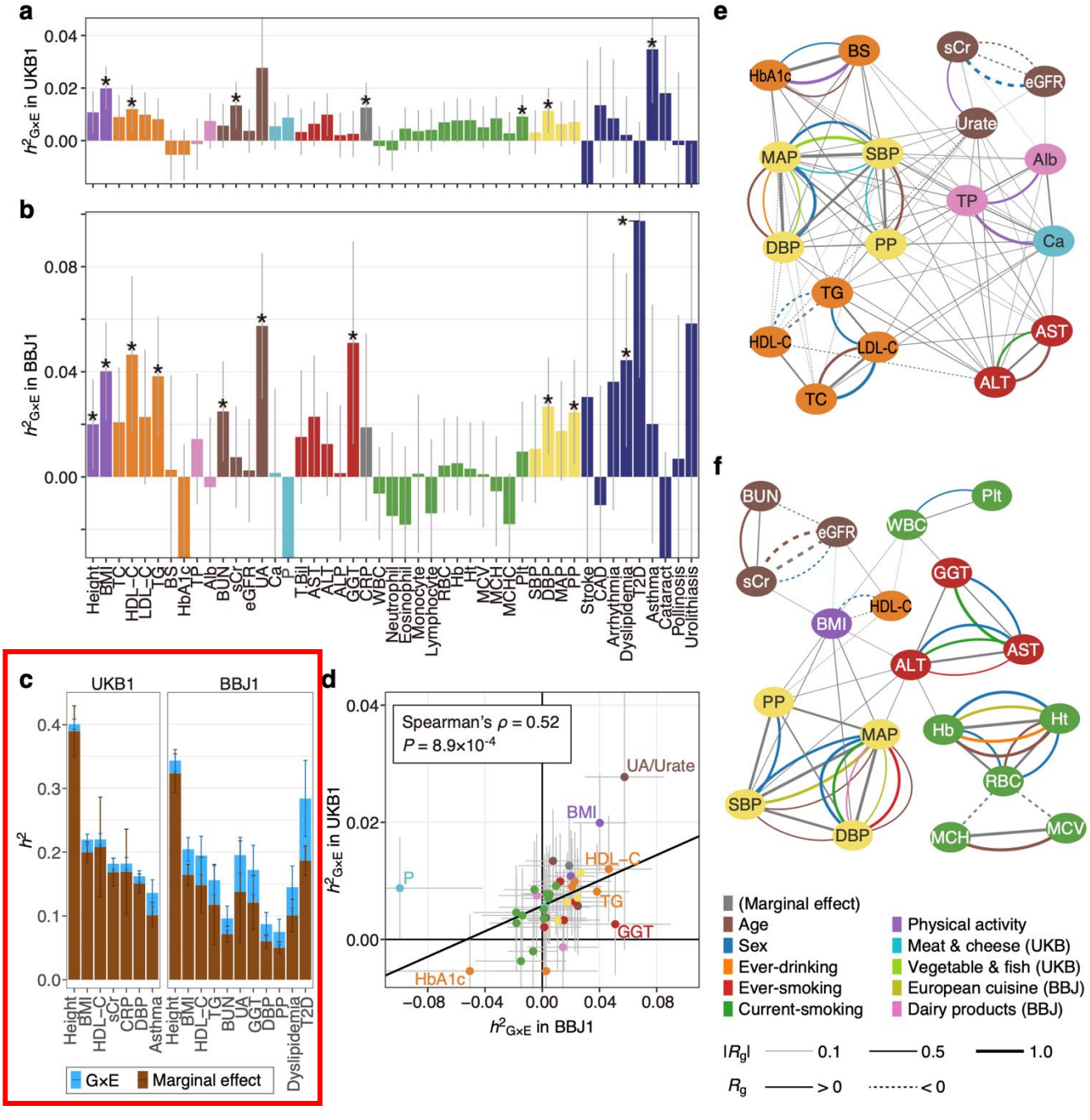


Figure 3, Genome-wide consistency of G×E interactions across populations

a, G×E heritability (h^2) in UKB1, aggregated across all environmental factors (**Methods**). Asterisks denote FDR < 0.05; error bars show 95% CI. **b**, same as **a**, for BBJ1. **c**, heritability of G×E interactions and marginal effects for traits with significantly positive G×E heritability. **d**, scatter plot of G×E heritability in UKB1 and BBJ1. Error bars represent 95% CI; regression line estimated by Deming regression to account for estimation errors in both *x*- and *y*-axes. **e**, significant genetic correlations (R_g, FDR < 0.05) of marginal effects (gray lines) and G×E interactions (colored lines), estimated in UKB1. The widths of lines represent the strength of genetic correlations; dashed lines represent negative genetic correlations. **f**, same as **e**, for BBJ1.

9) The PGS and cross-population portability fundamentally seems like a different question and scope than the rest of the paper detecting locus-specific GxE and biologic mechanisms. Could be a separate and equally impactful brief report. As is, it gets lost in the length of the results and is important to the field!

Response:

Thank you for kindly acknowledging the importance of our analyses on PGS and cross-population portability. Our manuscript aims to provide a comprehensive overview of G×E interactions at both the variant and genome-wide levels. This section was intended to complement the genome-wide evaluation of G×E interactions, alongside the preceding sections on G×E heritability and cross-trait G×E correlations. Therefore, we would like to retain this section in our manuscript, while we agree with you that making our manuscript more concise will further improve its readability. Following your comment, we have shortened the text that was not directly related to the reviewers' comments. However, in this revision, we have refrained from rigorously shortening the parts related to your and other reviewers' comments, as we believe it is important to maintain transparency in our responses. We will continue working on further improving the conciseness of the manuscript.

10)The single cell PP follow up result also seems specific to a condition and its own important discovery but if length not an issue okay to keep in the paper.

Response:

Thank you for kindly acknowledging the importance of our single-cell analyses. As with the PGS and cross-population portability analyses discussed in our response to Comment #9, this section is a part of the genome-wide evaluation of G×E interactions. Throughout the manuscript, we aimed to explore the biological mechanisms underlying G×E effects, and this analysis provides a representative example of how G×E interactions can be linked to specific cellular contexts. Therefore, we would like to retain this section as an illustrative case of biological insights on G×E interactions. We revised our manuscript to make it more concise and to reduce overall length, as noted in our response to Comment #9.

Response to the reviewer's Comments:

Reviewer #1:

My previous concerns have been fully addressed by the authors. I have only one minor comment remaining:

Response:

We sincerely thank you for taking your time to consider our manuscript and kindly providing insightful comments to improve the manuscript further.

Lines 485–498 – Reference for G×E-PGS:

In the section beginning "Consistent with these findings, PGS constructed from G×E interactions (G×E-PGS) significantly...", the authors note that elastic net was used for the G×E-PGS analyses. However, the model itself—comprising PGS_additive, PGS_G×E, and the interaction term PGS_G×E × E—is conceptually based on the G×Eprs framework. As such, the G×Eprs model should be properly referenced to acknowledge that this modeling approach is pre-existing and has already been validated. While the supplementary methods provide some explanation, the manuscript would benefit from a citation to an established reference, if available.

If, on the other hand, G×E-PGS is intended to represent a novel methodological contribution, this should be clearly stated in the manuscript. In that case, it would also be important to support the approach with simulations or additional analyses to demonstrate its validity and robustness.

Response:

Thank you for your comment. As you pointed out, we imported the concept of G×E-PGS from the G×Eprs paper that you kindly introduced. We completely agree with you that we should cite the paper in the main text to properly acknowledge it, and we have revised our manuscript accordingly.

Modification:

"Consistently, PGS constructed from G×E interactions (G×E-PGS)[40] significantly explained phenotypic variance in independent cohorts" (**Main, p.16, lines 368–369**).

"40. Jayasinghe, D. *et al.* Mitigating type 1 error inflation and power loss in GxE PRS: Genotype–environment interaction in polygenic risk score models. *Genet. Epidemiol.* **48**, 85–100 (2024)." (**Main references, p.24, lines 586–588**).

Reviewer #2:

I greatly appreciate the author's thorough response to reviewer comments. In nearly all areas, I am satisfied with their responses. I particularly appreciate removing some over-stating claims and incorporating additional cohorts and longitudinal data. I still have 1 major and 3 minor comments

Response:

We truly appreciate your careful review of our manuscript and the insightful comments you provided in both the previous and this round of revisions, which have substantially improved the manuscript.

Major comment

1. Mendelian randomization (MR) - While MR can be a powerful approach for assessing causal relationships, its assumptions are frequently violated, especially in studies of environmental exposures. In this context, the validity of the genetic instruments is particularly questionable due to the high likelihood of horizontal pleiotropy.

The authors rely solely on the inverse-variance weighted (IVW) method, which assumes that all genetic instruments are valid and that there is no directional pleiotropy. However, even a single pleiotropic variant can bias the results. No sensitivity analyses (e.g., MR-Egger, weighted median, or heterogeneity tests) were performed to assess pleiotropy or potential violations of MR assumptions.

Regarding the Natto/warfarin example, I believe temporal ordering alone provides sufficient evidence for the reverse causal relationship, and the MR analysis adds little value—particularly without robust validation of the instruments. Unless the authors can convincingly justify that the genetic instrument for Natto consumption is valid and not affected by horizontal pleiotropy, I recommend removing this MR analysis.

I also encourage the authors to consult the MR-STROBE guidelines and include a checklist if MR is retained in any part of the manuscript. The recent preprint by Sutton et al. (https://doi.org/10.1101/2025.06.26.25330002) provides an overview of the concerns in conducting MR with dietary exposures and reinforces the need for rigorous instrument validation.

Overall, I recommend removing the MR component from the manuscript unless a more in-depth analysis is performed, with appropriate sensitivity checks, justification for instrument validity, and MR STROBE checklist (https://www.strobe-mr.org/). An additional concern that I have with retaining it, is that this publication is already dense, and overly-detailed with a battery of tests. Any more detail in MR would further decrease the readability of the manuscript.

Response:

Thank you for your insightful comment. Further improving the readability of the manuscript was also emphasized by the editor, and we fully agree with your recommendation. We conducted a sensitivity analysis using the MR-Egger method, and the two significant associations identified by the IVW method (previously described in **Supplementary Note 8**)

were no longer significant. Considering this and your helpful suggestion, we have decided to remove the MR results to improve the clarity and focus of the manuscript.

Minor comments

1. In Line 462 you discuss the lack of correlation between heritability of vQTLs and GxE interactions is not necessarily surprising, given your GxE interaction analyses of only a small subset of exposures is expected to capture all the underlying variance, which could be due to many other GxE and GxG. I think your interpretation of this result is not accurate - that vQTL failed to fully capture GxE interactions. Please modify this.

Response:

Thank you for your insightful comment. We agree with you that the vQTL might be related to G×E interactions from environments not evaluated in our analysis. We have removed the sentence that “vQTL failed to fully capture G×E interactions” from both the main text and **Supplementary Note 9**.

Modification:

We also added the following sentence in **Supplementary Note 9**.

“On the other hand, it is also possible that vQTL might be related to G×E interactions from environments not evaluated in this study.” (**Supplementary Note 9**).

2. Although the results catalog is more comprehensive than prior studies, the manuscript still overstates its scope by referring to the analysis as "phenome-wide" or "exposome-wide" when in fact only a limited number of phenotypes and exposures were tested. I recommend softening these claims to better reflect the study's actual breadth.

Response:

Thank you for your comment. Following your suggestion, we removed the phrases “phenome-wide” and “exposome-wide” from our manuscript. We note that the phrase “phenome-wide” is now used exclusively to mention our phenome-wide association study (PheWAS) approach for evaluating the pleiotropy of G×E interactions across diseases.

3. The manuscript remains highly information-dense, to the point of being difficult to read. I encourage the authors to reduce the number of details in both the text and figures wherever possible. Streamlining the presentation would improve readability.

Response:

Thank you for your insightful comment. Further improving the readability was also emphasized by the editor, and we revised the manuscript to make it concise.

Reviewer #3:

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Response:

We sincerely thank you for kindly reviewing our manuscript and providing insightful comments that have substantially improved the manuscript.

Reviewer #4:

Thank you for careful consideration of the initial comments provided. The revisions strengthened aspects of the prior version that were concerns for the reviewers and I have no further concerns at this time.

Response:

We sincerely appreciate your time and expertise in reviewing the manuscript and thank you again for your many thoughtful comments during the revision.