nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Analysis of blood methylation quantitative trait loci in East Asians identifies ancestry-specific effects associated with complex trait variation
Corresponding author name(s): Andrew E. Teschendorff, Fan Liu, Sijia Wang

Reviewer Comments & Decisions:

Decision Letter, initial version:

4th Jan 2022

Dear Professor Wang,

Your Article entitled "Comprehensive mechanistic characterization of mQTLs in an Asian population" has now been seen by 3 referees, whose comments are attached. While they find your work of potential interest, they have raised serious concerns which in our view are sufficiently important that they preclude publication of the work in Nature Genetics, at least in its present form.

While the referees find your work of some interest, they raise concerns about the strength of the novel conclusions that can be drawn at this stage.

Briefly, one reviewer is positive and supportive of publication. The other two referees, however, while acknowledging the value of your East Asian ancestry cohort, identify substantial issues with your manuscript. Most notably, they have important questions regarding the computational methods used (fastQTLmapping, CellDMC, OpenCausal), technical aspects of the analysis (the number of independent mQTLs identified, definitions of backgrounds used in the enrichment analyses), as well as the biological findings (overall novelty, causality of methylation on traits).

Given the broad range of these criticisms, we concluded that a major revision would be required to address these comments; and even then, it remained unclear to us whether these critical referees would then be supportive of publication at Nature Genetics.

Should further data allow you to fully address these criticisms we would be willing to consider an appeal of our decision (unless, of course, something similar has by then been accepted at Nature Genetics or appeared elsewhere). This includes submission or publication of a portion of this work someplace else.

The required new experiments and data include, but are not limited to those detailed here. We hope

you understand that until we have read the revised manuscript in its entirety we cannot promise that it will be sent back for peer review.

If you are interested in attempting to revise this manuscript for submission to Nature Genetics in the future, please contact me to discuss a potential appeal.

I would also be happy to consult with our sister journal, Nature Communications, who would likely be willing to consider a less-extensive revision than would be required for consideration at Nature Genetics. This would present a more rapid path to publication. Please get in touch with me if you would like to pursue this option.

Otherwise, we hope that you find our referees' comments helpful when preparing your manuscript for resubmission elsewhere. Thank you for giving us the opportunity to consider your work.

Sincerely,

Michael Fletcher, PhD
Associate Editor, Nature Genetics

ORCID: 0000-0003-1589-7087

Referee expertise:

Referee #1: methylation, (epi)genetics/genomics, population health.

Referee #2: methylation/epigenetics, epidemiology.

Referee #3: methylation, statistical genetics.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
In this manuscript, Peng et al. presented their work on the characterization of methylation quantitative trait loci (mQTLs) in a Han Chinese (Asian) population, as well as investigated the cell type specific nature of the mQTLs and associated molecular mechanisms.

General comments
The authors set out strategic motivations for their work in the initial introduction:
1. Understand how epigenetically mediated genetic predisposition to disease could preferentially affect specific ethnicities
2. Establish If DNAm changes driven by SNPs display cell-type specificity
3. Investigate the role of SNPs in modulating chromatin accessibility
4. Previous mQTL studies have used the 450K beadchip, with lower resolution that the current 850K beadchip.

While the ambitions are reasonable, it was unclear whether the work presented achieves the ambitions.
1. The results seem to suggest that ~90% of EAS mQTLs replicate those In Europeans. Replication alone is somewhat incremental as a finding. The authors do not analyse for heterogeneity of effect between populations, and there is no analysis for the significance (or validity) of 'ethnic specific' mQTLs. Do the 'population specific effects replicate? What are the insights we glean, or is this just more of the same? At the very least, it might have been Interesting to see some ethnic specific GWAS signals evaluated through functional genomic strategies. At present this study could have been done in any population group.
2. Cell-type specificity. The CellDMC analysis is limited to SNPs that are associated with DNAm in the discovery cohort. Since this is a mixed cell population, it is anticipated that the variants identified will be biased towards those that have similar effects across white cell groups. SNPs with heterogenous (or opposite effects) will be masked and missed. Unsurprisingly the authors report that ~90% of SNPs have similar effects across cell subsets. What are the Insights from this analysis?
3. That mQTL SNPs impact and associate with chromatin accessibility has been previously shown, and I was unclear what the substantive new insights are.
4. The 850K array has been used in a number of mQTL studies, Peng et al are not the first and results are not compared to these other efforts. In any event: i. coverage remains low (<5%) even with the 850K array (that has been around for at least 5 years) with most of the genome not assessed, so the improvement could be seen as incremental; ii. what are the insights that are generated by use of the 850K array that were not seen with the 450K array (beyond a few more mQTLs)? What is the new insight? What do we learn form the newly identified mQTLs? iii. Since 95% of markers CpG markers are not assessed, have the authors considered fine-mapping by resequencing to generate more precise information on causal SNPs and CpGs?

As another general note, the authors state the work is important because East Asians are the 'largest ethnic group' (line 66). I am not sure that this is true, and in any event: i. does this single population cohort study represent the tremendous diversity of East Asia, and ii. as noted above, the study provides only limited cross-ethnic analyses. It is unclear what ethnic specific insights are generated if any, beyond showing that most associations are the same across ethnic groups.

Specific technical comments

The results obtained in the current study hinges heavily on their newly developed tool FastQTLmapping, which appears to be impressive in terms of compute time/resource needed, but is currently unpublished yet. I am not sure that the current version available on biorxiv is sufficiently detailed for me to be fully convinced that the results obtained are reliable, as compared to MatrixEQTL. For instance, there is only one figure comparing the computation and I/O time as well as peak memory consumption in the available draft (no table or Supp Materials). The author stated that in the presence of missing values, fastQTLmapping achieved results that were always closer to the exact results, but with no further details on the extent/type of missing values, nor the actual method of imputing the missing values.

The authors reported a total of 62.92M genome-wide significant mQTLs, including 56.29M cis-, 2.27M lcis- and 4.36M trans-mQTLs at Bonferroni adjusted p-values of Pcis <1.06×10-11, Plcis<2.86×10-12 and Ptrans < 8.16×10-15. As we know, there exists strong LD between SNPs, and also some level of correlation between methylation markers (but not as strong due to the sparse nature of the markers

on the array). It will be important to establish how many 'independent' associations these mQTLs represent.

The authors hypothesize that ethnic-specific mQTLs likely exist, in view of the small number of mQTLs identified in the East Asian population that were not reported in previous Caucasian studies. However, as mentioned in Discussion, the authors were not able to eliminate the contribution of other factors such as statistical power and differences in the versions of the methylation arrays (450K vs EPIC).
- It is critical that the authors demonstrate that the study is sufficiently powered to robustly identify these 'ethnic-specific'/novel loci.
- The authors demonstrated that 81% of ~16M mQTLs with genome-wide significance in the FHS study [Huan et al. 2019] were also significant in the current East Asian cohorts at FDR < 0.05. It will be informative to understand what is the proportion in Huan et al. that achieved statistical significance for the genome-wide significant mQTLs in the current study, restricted to methylation markers on the 450K array.
- Tha authors focuses the study on the ethnic population (East Asians) that it was conducted in, as well as the potential role of mQTLs in mediating disease risk. It will be interesting to evaluate if the SNPs (and possibly CpGs) from the proposed panel of 'ethnic-specific' mQTLs are enriched for disease/phenotype traits that are known to differ in risk/prevalence between East Asians and Caucasians.
On a related noted, it will be informative to consider sex-specific mQTLs.

The choice of samples for the East Asian validation dataset should be clarified. The samples were obtained from participants in two clinical trials of chiglitazar, which is tested for use in the treatment of Type 2 Diabetes. It is unclear from the current manuscript if blood was taken at baseline (before chiglitazar treatment) or after. However, it is most likely safe to assume that the participants are individuals suffering from T2D. We know that methylation levels can be modified by drug treatments, and also from previous studies that methylation profiles are altered in T2D patients, even before disease onset. The study will benefit greatly from a validation series that is population-based.

There are also some concerns on the technical methodology. Firstly, the authors have opted for a 2-step analysis strategy for mQTL mapping, including the excluding of outlying methylation values in the 2nd step. It is unclear why the authors opted for this 2-step strategy, and also the reason for excluding 'extreme' methylation values (defined as outside the range of mean±3SD). This methylation values may well be the most informative, and the authors should also assessed the impact of this filtering.

It is also stated that in the mQTL analysis, adjustments were made for 'bisulfite slide number' (please clarify what this refers to), batch (again, please clarify this), as well as the top 2 DNAm PCs. What is the rationale for adjusting for the top two DNAm PCs?

In addition, it was stated in the Methods that missing beta values were imputed by impute.knn. Whilst it is not uncommon to impute for missing values, it is worth noting that in a recent comparison of methylation data imputation performances across seven methods [Lena et al. 2020], it was concluded that impute.knn is not suitable for DNA methylation data imputation, and that in general, it will be prudent to accompany data imputation by sensitivity analyses.

The selection of control/background group is critical in enrichment analyses. For instance, for the enrichment analysis of mQTLs in 3D chromatin contacts, the authors defined the control groups as i)

random sampling of SNP-CpG pairs from all SNP-CpG combinations, regarded as genomic background and ii) SNP-CpG pairs with the same distance distribution as mQTL pairs (distance-matched SNP-CpG). This represents rather loose matching criteria, without taking into consideration other important factors such as MAF of SNP and variation of methylation level at CpG.

For the audience to better appreciate the mQTLs, it will be helpful for the authors to quantify the methylation effect size, and also with respect to whether the effect sizes were associated with stronger biological implication (e.g. association with gene expression and disease/phenotypic traits), as well as extent of reproducibility.

Finally, I believe that the study will be greatly strengthened by wet-lab experimental validation that could provide convincing evidence of the proposed causal role that mQTLs play, and/or the central role of transcription factors in bridging genetic variants and methylation levels.

Reviewer #2:
Remarks to the Author:
The authors are undertaking a worthwhile examination of the genetic variants that control DNA methylation in an East Asian population, to address concerns in the field over differences in this regulation by genetic race as well as differential genetic regulation by cell subtype. They demonstrate extensive overlap of mQTLs with a prior study in a European ancestry population, and using innovative cell specificity estimation demonstrate substantial overlap of these traits across cell types. They go on to describe the potential influence of chromatin architecture and develop a better understanding of how trans-mQTLs may be operating, and link their findings to two important phenotypes. These are significant and original findings with implications for downstream research efforts.

Title: Suggest adding that this is "in blood", eg. "Comprehensive mechanistic characterization of mQTLs in blood in an Asian population, " given the cell type specific nature identified. It is likely that there may be further differences in other tissues that were not examined in this study.

Methods: The authors have applied state of the art methodologies in the determination of mQTLs, including developing a novel C++ based algorithm to enhance the speed of detection of those mQTLs and reduce computing burden. The application of the CellDMC method to estimate cell type specificity of the mQTLs is also highly innovative, and provides important additional information regarding cellular specificity of this genetic regulation. All statistical methods and inference appear appropriate and robust, and the incorporation of a number of large, publicly available datasets adds to the value of these findings and their interpretation.

The examination of FOSL1 and NFKB1 hotspots add to the understanding of trans-mQTLs and disease process. It would be helpful, though, to have a better understanding of why these two hotspots, of the 16 top 1% hotspots were chosen for further dissection. Were the other hotspots not in disease associated regions?

Discussion: In the examination of mQTLs and hotspots, the authors performed a mendelian randomization and concluded that the methylation at these regions does appear to be in the causal path of the outcomes examined (eosinophilia, obesity). This is in contrast to prior work on mQTLs (Min et al, Nat Genet 2021) which concluded that in most cases methylation was not causally mediating a

variety of traits, including blood specific traits. The authors should discuss their findings in light of these results.

Reviewer #3:
Remarks to the Author:

A. Summary of the key results

The authors describe the largest mQTL-mapping study in a Han Chinese population (n=3523) using DNA methylation measured in whole blood. The find over 80% of mQTLs in common with a similarly-sized white population (FHS, n=4170) and replicate 87% in smaller Han Chinese population (n=798). They apply CellDMC to their whole blood data to identify cell-type specific mQTLs and estimate that <10% of mQTLs are cell-type specific. They confirm the importance of transcription factors to the functional roles of trans-mQTLs and explore roles for DNA methylation in mediating the effects of trans-mQTL 'hot spots' on eosinophilia, ulcerative colitis and body mass index.

B. Originality and significance: if not novel, please include reference

This is the first significant mQTL-mapping study in an Asian population. To the reviewers knowledge, the largest previous Asian mQTL study included Chinese (n = 93), Indians (n = 83) and Malays (n = 78):

Kassam, I., et al (2021). Genome-wide identification of cis DNA methylation quantitative trait loci in three Southeast Asian Populations. Human molecular genetics, 30(7), 603–618.

Although cell-type specific mQTLs are reported, these were estimated from whole blood DNA methylation using the CellDMC software tool. Methods like CellDMC are still relatively new and untested. Preliminary evaluations and the validation reported in this manuscript indicate that any reported cell-type specific associations should be considered highly speculative. These cell-specific results should not therefore be considered a significant contribution to the literature.

The manuscript concludes with the claim that the described mQTL datbase is "an invaluable resource for understanding the genetic and epigenetic variations in disease predisposition between ethnic groups." Although this is likely true, the analyses of the manuscript mainly focus on mQTLs in common with previous European studies (e.g. trans-mQTL hotspot relevance to disease). I was expecting the study to focus on Asian-specific mQTLs and their potential role in diseases with higher prevalence in Asian populations.

The authors note that a variety of factors other than ethnicity could explain differences between their study and previous non-Asian studies (Line 414 "e.g., differences in power or Illumina beadarray version"). Although this is true and lack of a significant p-value should not be used conclude absence of an association, it still possible to compare mQTL effects between studies and note where effects are significantly different.

C. Data & methodology: validity of approach, quality of data, quality of presentation

The mQTL analysis appears to have been sound.

The cell-count specific mQTL analyses are highly speculative because, as noted earlier, the methods are still new and relatively untested and performance evaluations indicate high error rates. The manuscript should be more clear in showing how, although there is some evidence of validation, the validation is extremely limited and shows much higher error rates than we'd expect if the analyses has been performed in purified cell-type populations.

It should be expected that more cell-type specific associations should be observed in the more abundant cell types as cell-type specific signal in the bulk tissue data will be stronger. For some reason the authors use this observation to conclude that the more abundant cell types are more 'dominant in blood' (lines 165-168 and lines 171-173). Besides the uncertain meaning of 'dominance' in this case, this limitation of the data should not be used to draw biological/functional conclusions.

The analysis of chromatin accessibility requires rationale for what seem to be arbitrary decisions:
- Are models 1 and 2 the only possible models? How were they selected? The criteria for each appears to be quite specific.
- What proportion of the the mQTL associations should we expect to explain based on models M1 and M2. The analyses suggest that models M1 and M2 "explain 40% of mQTLs". Is this more than expected? Should we be proposing additional models to explain the remaining 60%?

The "OpenCausal tool" is applied with little explanation or rationale. The text should include a short introduction to what the tool is, how it assesses causal relationships and the limitations of those assessments.

The manuscript claims to provide evidence for "DNAm levels at NFKB1 trans-mQTLs being causal mediators for BMI, as opposed to being a consequence of BMI" (Lines 371-372). First, I think the statement should refer to DNAm levels at mCpGs of the NFKB1 trans-mQTLs. Secondly, and more importantly, this finding appears to contradict an extensive literature on DNA methylation in blood and BMI, including the Wahl et al and Mendelson et al (Plos Med 2017) studies, which find almost no evidence for a causal effect on BMI. More generally, Min et al (2021) report, based on a much larger sample size (n=30K), very little evidence for a causal effect of DNA methylation on any phenotype. The authors should more carefully investigate these appearent disagreements with previous studies. It isn't sufficient to just note that Min et al "did not specifically focus on trans-mCpGs co-localizing with TF-binding." The Min et al study was genome-wide and better powered, so it should have identified at least as many causal relationships.

A prominant claim in the paper is that clusters of trans-mQTLs tend to coincide with clusters of transcription factor binding sites. In the text, the authors confusingly refer to this as trans-mQTLs being "surrounded by TFs". The supplementary methods defines this as being located within 1Mbp of a predicted transcription factor binding site. If this indeed the definition, then the text should just clearly state this simple definition rather than leave it buried in the supplementary materials. I'm not sure that this definition makes any sense. How likely is it that a genetic variant will influence the binding of a transcription factor 1Mbp away? I would have expected a distance with a much smaller distance.

One piece of the evidence supporting this claim is the scatterplot in Figure 4b that is claimed to show

a correlation between the number of trans-mQTLs and the number of transcription factor binding sites on the same chromosome. The correlation between the two appears to be driven mainly by chr19. What is the correlation with and without chr19?

D. Appropriate use of statistics and treatment of uncertainties

The authors report 62.92M mQTLs. It is standard to also report the number of independent mQTLs.

In many places p-values well below the precision of floating point calculations are reported (e.g. p < 1.00×10^-323). Values this small are meaningless and should be replaced with a value that better represents the capabilities of the computers used for analysis (e.g. a typical recommendation is p < 2.22x10^-16). Statistical strength of associations with extremely low p-values is better expressed with summary statistics such as effect sizes and confidence intervals.

Many enrichment analyses are reported, in most cases using the hypergeometric test. Authors should take care to correctly specify the universe/background in these tests. In most enrichment tests, the text does not clearly indicate how the universe/background was defined. For example, line 124-125 says that "mSNPs were enriched in genomic functional regions such as promoters and exons, and this pattern was more pronounced for trans-mSNPs than cis-mSNPs". It is unclear whether or not this enrichment accounted for the fact that DNA methylation measurements on the Illumina Beadchips are highly enriched in promoters and exons.

E. Conclusions: robustness, validity, reliability

Conclusions about cell-type specificity should be strongly qualified in light of the methods used and validation findings.

F. Suggested improvements: experiments, data for possible revision

Overall, the manuscript text needs to be revised to use technical terms correctly and precisely and to simplify text that is unnecessarily complex.

The chromatin analysis needs to be better explained and justified.

To capitalise in the major contribution of this study, mQTLs in an Asian population, the authors should make an effort to identify ethnicity-specific mQTLs and investigate their potential role in ethnicity-specific disease.

G. References: appropriate credit to previous work?

Shoud cite the largest previous Asian mQTL study included Chinese (n = 93), Indians (n = 83) and Malays (n = 78) and compare findings:

Kassam, I., et al (2021). Genome-wide identification of cis DNA methylation quantitative trait loci in

three Southeast Asian Populations. Human molecular genetics, 30(7), 603–618.

It is somewhat unexpected that findings are not compared to the largest mQTL study carried out so far (n=30K, Min et al. 2021).

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

A lot of the text in the results section is unnecessarily complex. For example, consider the following sentence on lines 100-101:
"The mQTL SNPs (mSNPs) covered more than 2/3 of tested SNPs (5.56M), while mQTL CpGs (mCpGs) covered 1/3 of tested CpGs (284,128)."
Here is a simpler version:
"Two-thirds of the tested SNPs (5.56M) were associated with DNA methylation, while one-third of tested CpG sites (284,128) were associated with genetic variation."
This example highlights two causes of unnecessary complexity that appear repeatedly throughout the results section. The first cause is the misuse of terms already well-defined in the literature. In this example, term "mQTL SNP" is redundant because an mQTL is by definition a SNP, a SNP that is associated with DNA methylation at a CpG site. Thus, "mSNP" is an unnecessary definition because mQTL and mSNP are equivalent. The second problem is unusual choices of words and phrases. The term "covered" here is confusing because it suggests a more complex relationship between mQTLs and SNPs than that mQTLs are simply a specific subset of SNPs that are associated with DNA methylation. The results section needs to be revised to simplify the text and ensure correct use of defined terms.

Another important example is references to trans-mQTL "hot-spots" which are very simply defined as genomic loci containing a large number of trans-mQTLs. However, confusion is caused by reference to a "trans-mQTL network" (e.g. Line 255) which is never defined and to "unlinked trans-mCpGs" whose vague definition is buried in the supplementary materials. There CpG sites are mysteriously "clumped" in 500Kbp windows to "exclude linkage among adjacent CpGs". The term "linked" here is non-standard and actually incorrect because it refers to "linkage disequillibrium" (LD). LD is about genetic variation, not DNA methylation variation. It is more typical to refer to an "index" CpG site which representats a cluster of strongly correlated CpG sites. The "hotness index" would then be defined for a cluster of trans-mQTLs in linkage disequillibrium as the number of associations of these mQTLs with trans index CpG sites.

Below are other examples:

Line 92 algorithm, fastQTLmapping was up to 4 and 11 times faster in the single-thread and 32 CPU threads
"up to" isn't very meaningful, summarize with ranges or averages

Line 119 allele frequency differences for pan-ethnic mSNPs were significantly smaller when compared to the

I assume the allele frequency differences referred to are between FHS and EAS.

Line 120 18.91% mSNPs that were only significant in FHS

This is an unnecessary use of jargon. Better to say "mSNPs that were only observed in FHS" something like that.

Line 194 cis-mQTL pairs

I suspect that this refers to pairs of associated CpG sites and cis-mQTLs. However, this term is not correct.

Line 256 With hotness-index
Line 257 increasing, the proportion of mSNPs in hotspots surrounded by TFs was monotonically increasing

Simpler to write: As the hotness-index increases, the proportion of trans-mQTLs within 1Mbp of a transcription factor binding site increases monotonically.

Figure 2a doesn't add up. The myeloid/lymphocyte analysis reports about 2E7 mQTLs in lymphocytes but about twice that number in specific lymphocyte cell types. By contrast 6E7 mQTLs are reported in myeloid cells but about half that number in specific myeloid cell types.

Line 310 The majority (141, 60.8%) of the 232
Line 311 mQTLs were detected exclusively in the myeloid lineage (Fig. 5b&c, Fig. S31b).

Okay, but how strong is the evidence that they do not occur in lymphocytes? There appears to be something wrong with Figure 5b. There are 232 mQTLs, 141 are detected in myeloid cells but none in lymphocytes?

Line 312 Atlas31, we found the 232 mCpGs to be mainly enriched in immune system disorders (P.bfadjust =

I assume that "P.bfadjust" is just refers to a Bonferroni adjusted p-value.

In Figure 5f, "Kim data" should be replaced with a more formal citation of the dataset.

Line 356 Another important trans-mQTL hotspot was driven by a GWAS SNP associated with ulcerative colitis
Line 357 (UC) and linked in-cis with the transcription factor NFKB1

What does it mean for a hotspot to be 'driven by' a specific SNP?

Line 358 al (2017)1, and thus validating this NFKB1 trans-mQTL network in an EAS population

Unsure where "trans-mQTL network" is defined.

Line 364 that a list of 364 CpGs known to be associated with BMI (as derived from Wahl et al (2019)64 and
Line 365 other studies), did so also in our Asian cohort and with the same directionality of DNAm

change

Simpler to say "that published associations of 364 CpG sites with BMI (Wahl et al 2019) were replicated in our Asian cohort"

Line 419 associated with environmental factors.

I don't understand this sentence.

Line 792 g, Enrichment of mQTL pairs in functional elements.
Line 794 Heatmap shows the fold changes (see Methods) of SNP-CpG pairs in all combinations of functional
Line 795 categories.

Fold changes with respect to what?

Decision Letter, Appeal:

29th Jun 2022

Dear Sijia,

Thank you for your message of 29th Jun 2022, asking us to reconsider our decision on your manuscript "Comprehensive mechanistic characterization of mQTLs in an East Asian population". I have now discussed the points of your letter with my colleagues, and we think that your appeal on our previous decision has addressed the major points highlighted to our satisfcation. We therefore invite you to submit the revised manuscript for peer review by the original referees.

When preparing a revision, please ensure that it fully complies with our editorial requirements for format and style; details can be found in the Guide to Authors on our website (http://www.nature.com/ng/).

Please be sure that your manuscript is accompanied by a separate letter detailing the changes you have made and your response to the points raised. At this stage we will need you to upload:
1) a copy of the manuscript in MS Word .docx format.
2) The Editorial Policy Checklist:
https://www.nature.com/documents/nr-editorial-policy-checklist.pdf
3) The Reporting Summary:
https://www.nature.com/documents/nr-reporting-summary.pdf
(Here you can read about the role of the Reporting Summary in reproducible science:
https://www.nature.com/news/announcement-towards-greater-reproducibility-for-life-sciences-research-in-nature-1.22062)

Please use the link below to be taken directly to the site and view and revise your manuscript:

[redacted]

With kind wishes,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Author Rebuttal to Initial comments

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
In this manuscript, Peng et al. presented their work on the characterization of methylation quantitative trait loci (mQTLs) in a Han Chinese (Asian) population, as well as investigated the cell type specific nature of the mQTLs and associated molecular mechanisms.

General comments
The authors set out strategic motivations for their work in the initial introduction:
1. Understand how epigenetically mediated genetic predisposition to disease could preferentially affect specific ethnicities
2. Establish If DNAm changes driven by SNPs display cell-type specificity
3. Investigate the role of SNPs in modulating chromatin accessibility
4. Previous mQTL studies have used the 450K beadchip, with lower resolution that the current 850K beadchip.

While the ambitions are reasonable, it was unclear whether the work presented achieves the ambitions.
1. The results seem to suggest that ~90% of EAS mQTLs replicate those In Europeans. Replication alone is somewhat incremental as a finding. The authors do not analyse for heterogeneity of effect between populations, and there is no analysis for the significance (or validity) of 'ethnic specific' mQTLs. Do the 'population specific effects replicate? What are the insights we glean, or is this just more of the same? At the very least, it might have been Interesting to see some ethnic specific GWAS signals evaluated through functional genomic strategies. At present this study could have been done in any population group.
Response: Thank you for these valuable suggestions. Following these suggestions, we performed a cross-ethnic comparison using the recent large-scale meta-analysis of European cohorts from the Genetics of DNA Methylation Consortium (GoDMC). These are the main findings: 1) Among the 2.65 million NSPT mQTLs, the majority of them (2.41 million or 91%) were also study-wide significant mQTLs in GoDMC. The fact that the majority of the mQTLs

are not population specific holds true regardless of the significance threshold used. 2) The remaining 9% (238K) mQTLs, regarded as East Asian specific mQTLs in NSPT, could be replicated very well in another East Asian cohort CAS (99.6% replicated). 3) For a mQTL in East Asians, the likelihood of it being also a study-wide significant mQTL in Europeans heavily depended on its MAF in the Europeans, and vice versa. 4) The enrichment analysis of East Asian specific mQTLs in GWAS catalog implied a pronounced enrichment in diseases/traits with known prevalence differences between populations (e.g., attention deficit hyperactivity disorder[4], bipolar disorder[5], pancreatic cancer[6], and trans fatty acid levels[7,8]).

We added these new analyses in Results on Pages 8~9 '**East-Asian specific mQTLs**' and also in Discussion on Page 27~28.

2. Cell-type specificity. The CellDMC analysis is limited to SNPs that are associated with DNAm in the discovery cohort. Since this is a mixed cell population, it is anticipated that the variants identified will be biased towards those that have similar effects across white cell groups. SNPs with heterogenous (or opposite effects) will be masked and missed. Unsurprisingly the authors report that ~90% of SNPs have similar effects across cell subsets. What are the Insights from this analysis?

Response: The reviewer has raised an excellent point. There are two reasons why CellDMC was run on mQTLs from the discovery phase. First, the presence of an interaction term in the CellDMC model means that this model is not amenable to analysis with MatrixEQTL or our own FastQTLmapping technique. Thus, the computational burden to run trillions of regressions with an interaction term included is several orders of magnitude higher than running trillions of ordinary regressions. Second, it is natural to attempt identifying cell-type specific mQTLs among the discovered mQTLs, since it is important to establish in which cell-type(s) a discovered mQTL is present in. The reviewer is concerned that our strategy (i) misses many non-mQTLs that are cell-type specific mQTLs, and (ii) that the mQTL-selection step favours mQTLs that occur unidirectionally in all underlying cell-types. We take the view that these are very minor limitations, and that the strategy followed by us is the best possible one we can pursue at present. First, let us note that all current evidence points to most mQTLs being cell-type independent. There are by now at least 4 independent studies supporting this: (i) the BLUEPRINT paper[9] profiled DNAm in 3 sorted blood cell subtype populations concluding that at least 75% of mQTLs were shared by the 3 blood cell subtypes, (ii) by re-analyzing the BLUEPRINT data using an independent unsupervised tensorICA method we obtained a similar lower bound of 75% on the fraction of cell-type independent mQTLs, (iii) the recent study by Hawe et al (2022)[10] also concluded that even between very distinct cell-types such as immune and fat cells, the great majority of mQTLs are shared, (iv) our own estimate derived from simulation models, and as presented in this work, indicates that approximately 90% of mQTLs are shared between blood cell subtypes. This estimate is consistent with the lower bound estimate from BLUEPRINT (recall that BLUEPRINT only profiled about 200 samples per cell-type, and thus is underpowered to detect mQTLs present in all 3 cell-types). Thus, the great majority of mQTLs

will be shared by blood cell subtypes. Hence the key question is whether we are identifying the ~10% of cell-type specific mQTLs, and the answer here is yes. For instance, among all mQTLs ($P<1 \times 10^{-14}$), we find that approximately 9% are either clearly myeloid or clearly lymphoid specific. In other words, we are identifying cell-type specific mQTLs at the proportion we would have expected based on independent studies and analyses. In order to make this clear, we have now added a new panel-f to Fig.3, to display an example of a myeloid and lymphoid specific mQTL, to make it clear to this reviewer that the algorithm is finding cell-type specific mQTLs. For convenience we display the new panel Fig.3f below:

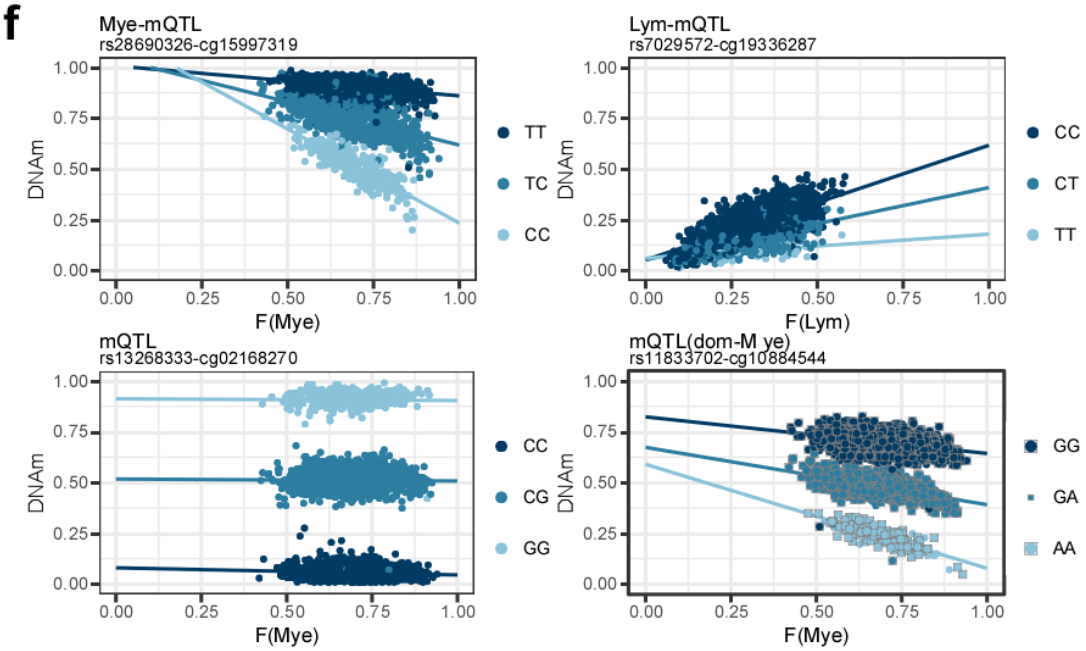


Figure 3f Scatterplots of DNAm (y-axis) vs cell-type fraction F (x-axis) for 4 mQTLs with samples colored by genotype.
The top two mQTLs are examples of a myeloid and lymphoid-specific mQTL, with the x-axis labeling the myeloid and lymphoid fraction, respectively. The bottom two mQTLs are examples of two cell-lineage independent mQTLs, with the left mQTL being equally dominant in myeloid and lymphoid subsets and the right mQTL being more dominant in the myeloid subset.

Second, it is important to stress that pre-selecting mQTLs based on P-values does not equate to selecting mQTLs based on effect-size. Whilst we agree that mQTLs that are unidirectionally present in all underlying cell-types may on average have larger effect sizes, this does not necessarily translate into a more significant P-value within the context of our interaction model. Indeed, it is worth noting that over 97% of mQTLs at $P<1◊10^{-14}$ display effect sizes (average DNAm change per allele copy) < 0.1.

The fact that the great majority of mQTLs display such low effect sizes is a very interesting

point, and although peripheral to the reviewer's point, is nevertheless of great importance to discuss, in order for the reviewer to appreciate the complexity of the mQTL effect size. In theory, if we assume that a given SNP induces a binary DNAm change at a given CpG within one cell, and if we assume that it causes this same effect in ALL cells from a GIVEN cell-type, then the effect size in that cell-type would be close to 1 (i.e. average DNAm difference would be close to maximal). If in addition we assume that the effect is unidirectionally present in all underlying cell-types within the tissue, then the effect size would still be close to 1. However, in reality we don't see this, partly due to technical limitations of the assay, which means that the observed maximum effect sizes are typically on the order of 0.8 to 0.9. Analyses from many mQTL studies (including ours) clearly demonstrate that mQTLs with such very large effect sizes are very uncommon, i.e. the great majority of mQTLs are of small effect size. This would suggest that either (i) many mQTLs are only present in subsets of cell-types, or (ii) that the mQTL effect within a given cell-type is heterogeneous, i.e. not all cells within a given cell-type exhibit the same binary DNAm change, or alternatively a mixed combination of scenarios (i)+(ii). To conclusively address this complexity will obviously require matched SNP and single cell DNAm data, which is clearly beyond the scope of this work. Given that there at least 4 independent studies and analyses suggesting that most mQTLs are shared between blood cell-subtypes, the dearth of large effect size mQTLs would suggest that scenario (ii) is very common.

3. That mQTL SNPs impact and associate with chromatin accessibility has been previously shown, and I was unclear what the substantive new insights are.
Response: Thank you for bringing up this issue. The new insights here can be concluded into two points: 1) Although mQTL SNPs impact and associate with chromatin accessibility has been shown, previous studies mainly discussed the influence of SNPs on chromatin, leaving the relationship between SNPs and CpGs not explored. Here in our work, we not only studied the influence of mQTL SNPs on chromatin, but also extended the mechanism to CpG end and explained the associations between SNPs and CpGs with 3D chromatin interaction. 2) Although previous studies have discussed the mechanism of mQTL by SNP-CpG enrichment in chromatin interaction regions[10], they did not clearly quantify the proportion of mQTLs explained by their mechanism. In our study, we proposed two possible regulatory mechanisms based on our knowledge in mQTL regulation, and systematically calculated the proportion of mQTLs explained by each mechanism. Collectively, the exploration of mQTL regulation in our work has its own novelty.
We have rewritten some sentences to clarify the significance of this analysis in Results on Page 14~15 and also in Discussion on Page 28.

4. The 850K array has been used in a number of mQTL studies, Peng et al are not the first and results are not compared to these other efforts. In any event: i. coverage remains low (<5%) even with the 850K array (that has been around for at least 5 years) with most of the genome not assessed, so the improvement could be seen as incremental; ii. what are the insights that are generated by use of the 850K array that were not seen with the 450K array (beyond a few more

mQTLs)? What is the new insight? What do we learn form the newly identified mQTLs? iii. Since 95% of markers CpG markers are not assessed, have the authors considered fine-mapping by resequencing to generate more precise information on causal SNPs and CpGs?
Response: We agree that this is not the first mQTL study about 850K, and we have weakened the relevant claim in the manuscript. In particular, we deleted the relevant sentence in Introduction, while only mentioned in Discussion that 'the use of 850K beadarray allowed us to more than double the numbers of mQTLs using the 450K array in studies with comparable sample sizes, providing a comprehensive picture of the mQTL landscape in East Asians'.

5. As another general note, the authors state the work is important because East Asians are the 'largest ethnic group' (line 66). I am not sure that this is true, and in any event: i. does this single population cohort study represent the tremendous diversity of East Asia, and ii. as noted above, the study provides only limited cross-ethnic analyses. It is unclear what ethnic specific insights are generated if any, beyond showing that most associations are the same across ethnic groups.
Response: Thank you for these suggestions. We deleted the 'largest ethnic group' description in Introduction.
For i), we believe that the use of "East Asian" is appropriate, given the context of the study, particularly after more cross-ethnic analyses have now been added. The samples collected in our cohort could represent the genetic background of East Asians, when comparing to the other world populations (e.g. European, African). As shown in Figure R1, our samples aggregate well with the other East Asian samples (EAS_CHB, EAS_CHS, EAS_JPT) from 1000 genomes project.
For ii), we agree that more cross-ethnic analyses should be added, and we have done so in the revised manuscript. The details have already been descripted in the response to comment 1.

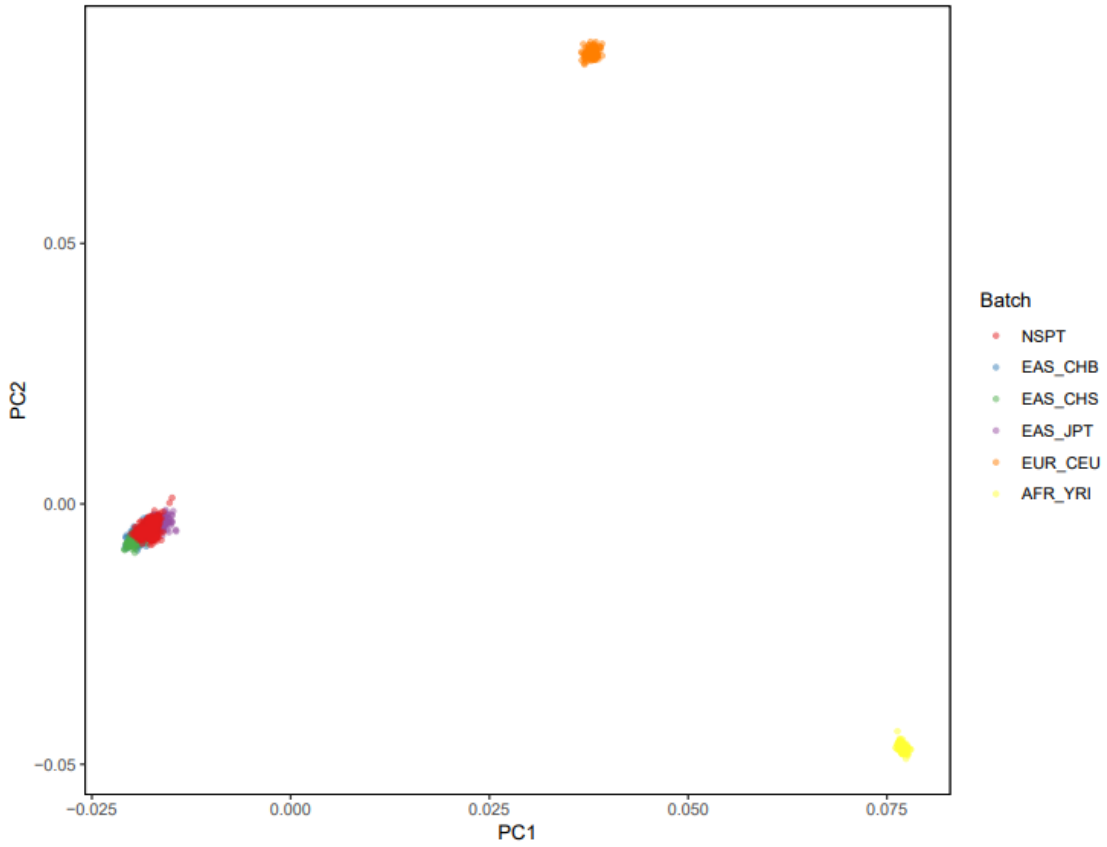


Figure R1 Population structure (Principal components) of our samples (NSPT) and samples from five other populations (AFR_YRI, EUR_CEU, EAS_CHB, EAS_CHS and EAS_JPT) from 1000 Genomes Project.
EAS_CHB: Han Chinese in Beijing, China
EAS_CHS: Han Chinese South
EAS_JPT: Japanese in Tokyo, Japan
AFR_YRI: Yoruba in Ibadan, Nigeria
EUR_CEU: Utah residents (CEPH) with Northern and Western European ancestry

Specific technical comments

6. The results obtained in the current study hinges heavily on their newly developed tool FastQTLmapping, which appears to be impressive in terms of compute time/resource needed, but is currently unpublished yet. I am not sure that the current version available on biorxiv is sufficiently detailed for me to be fully convinced that the results obtained are reliable, as compared to MatrixEQTL. For instance, there is only one figure comparing the computation and I/O time as well as peak memory consumption in the available draft (no table or Supp Materials). The author stated that in the presence of missing values, fastQTLmapping achieved results that

were always closer to the exact results, but with no further details on the extent/type of missing values, nor the actual method of imputing the missing values.

Response: Thank you for these suggestions. In this study, we preliminarily applied fastQTLmapping to screen significant mQTLs (P-value < 1×10^{-10}) utilizing its fast calculation. For the screened mQTLs, we then carried out these association independently in R. We found that the results from R were always consistent with those from fastQTLmapping, and the results reported in this study were based on the results from R. We briefly summarized the rationale and performances of fastQTLmapping below.

We developed a freely available C++ software package fastQTLmapping for fast mQTL analysis, which is applicable not only to all types of QTL-like analysis, but also to any forms of correlation or regression analysis between two extraordinarily large matrices. The package is released under GPL license and can be downloaded from https://github.com/TianTTL/fastQTLmapping. Figure R2 illustrates the general flow design of fastQTLmapping.

By utilizing the Level-3 BLAS math library and the OpenMP parallel computing framework, fastQTLmapping achieves extremely fast operation efficiency and stable memory consumption. We compared the performance of fastQTLmapping and MatrixEQTL. For a fair comparison, we made a parallel version of MatrixEQTL using R packages 'doParallel', linked MKL to R environment, and manually split data to feed MatrixEQTL to achieve its optimal performance. We randomly generated three sets of test data containing 10^9, 10^{10}, 10^{11} calculation tasks. We found that the computation and I/O of fastQTLmapping was 3.9-5.0 times and 5.4-11.7 times faster than MatrixEQTL respectively. The peak memory consumption of fastQTLmapping (1.3-14.5 GB) was much smaller than that of MatrixEQTL (7.1–78.8 GB) (Figure R3).

To deal with the potential errors occurring in covariate correction and missing value filling, fastQTLmapping is designed as a three-step procedure: first, using an optimized computational process to quickly obtain approximate xQTL results, then filtering candidate xQTLs with a relaxed significance threshold (by default, 100 times larger than the user-set threshold), and finally exhaustively performing multiple linear regression analysis on the candidate xQTLs. To test whether fastQTLmapping can reliably control the computational error, we randomly constructed 10 sets of test data with gradient missing rate (1% 2% 3% 4% 5% 6% 7% 8% 9% 10%), each containing 1 million association tasks. We applied MatrixEQTL, fastQTLmapping and R under the default parameters to carry out these calculations respectively. When taking the results from R as the golden standard, we found that the results of fastQTLmapping were always consistent with that from R. In contrast, the results of MatrixEQTL were less stable, and the error level was not related to the missing rate, but positively correlated with the significance level of the results (Figure R4).

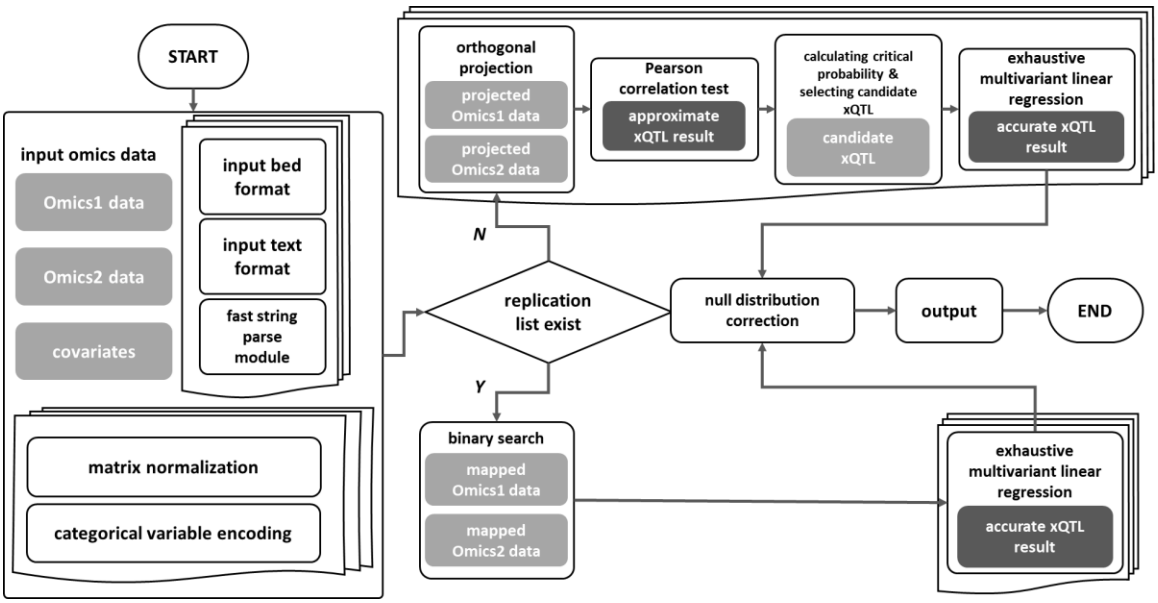


Figure R2. Flowchart of fastQTLmapping. Multiple documents symbols represent multiple tasks of parallel computing.

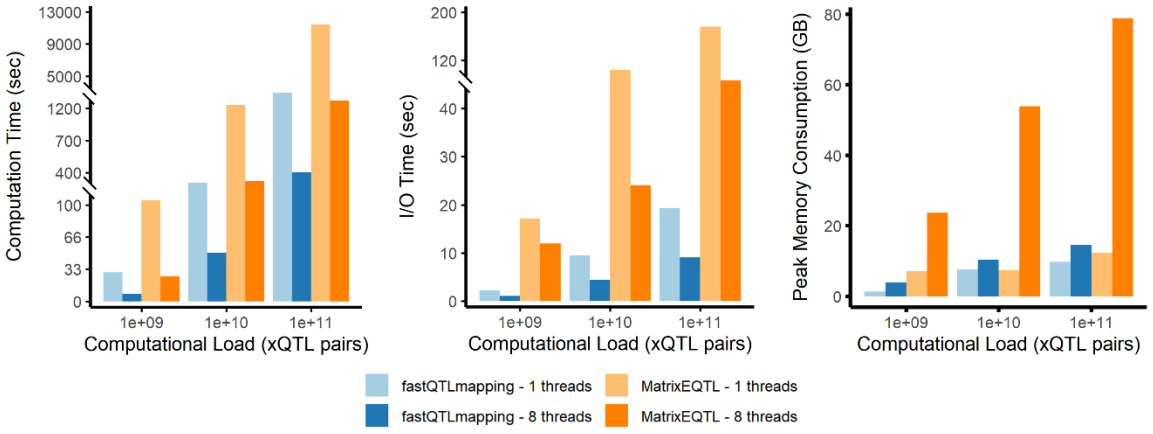


Figure R3. Performance of fastQTLmapping and MatrixEQTL under various settings.

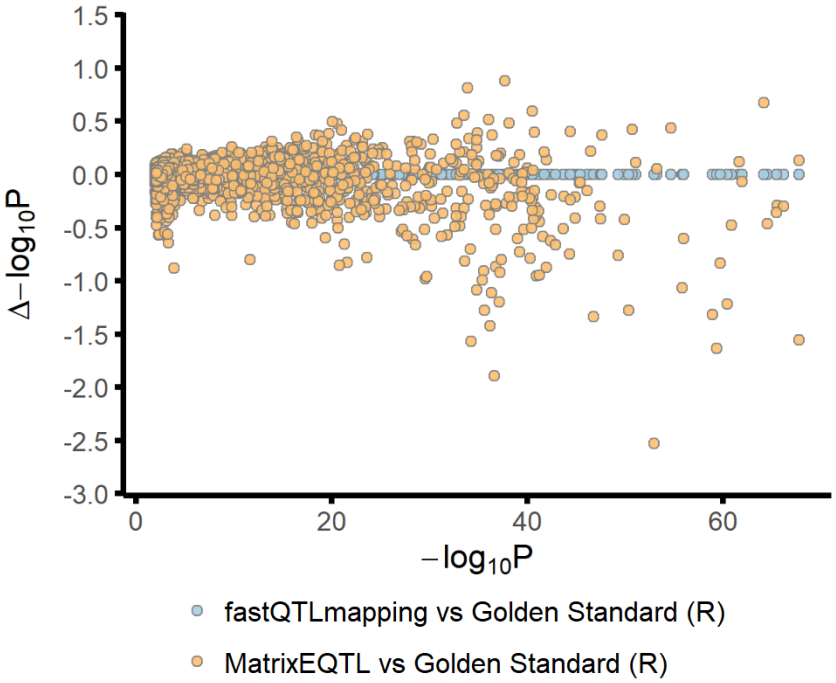


Figure R4. Error distribution of the results of fastQTLmapping and MatrixEQTL when analyzing omics data with missing values.

7. The authors reported a total of 62.92M genome-wide significant mQTLs, including 56.29M cis-, 2.27M lcis- and 4.36M trans-mQTLs at Bonferroni adjusted p-values of Pcis <1.06×10-11, Plcis<2.86×10-12 and Ptrans < 8.16×10-15. As we know, there exists strong LD between SNPs, and also some level of correlation between methylation markers (but not as strong due to the sparse nature of the markers on the array). It will be important to establish how many 'independent' associations these mQTLs represent.
Response: Thank you. We added the number of independent mQTLs in results. We found 56.29M cis-, 2.27M lcis- and 4.36M trans-mQTLs in whole genome. After pruning redundant SNPs in each category by limiting LD to r2<0.2, there remained 1.75M independent cis-, 52.25K lcis- and 111.63K trans-mQTLs. These results have been added in Results on Page 5.

8. The authors hypothesize that ethnic-specific mQTLs likely exist, in view of the small number of mQTLs identified in the East Asian population that were not reported in previous Caucasian studies. However, as mentioned in Discussion, the authors were not able to eliminate the contribution of other factors such as statistical power and differences in the versions of the methylation arrays (450K vs EPIC).
8.1 - It is critical that the authors demonstrate that the study is sufficiently powered to robustly identify these 'ethnic-specific'/novel loci.
Response: Thank you for this suggestion. The power of detecting population-specific mQTLs is

related to allele frequencies of each mQTL, its effect size, and sample sizes of the populations. For the mQTLs detected in East Asians but not replicated in Europeans, we estimated the power of detecting them in our cohort. We found that we have sufficient power to detect mQTLs at minor allele frequency of 0.01 when DNA methylation variance explained is larger than 2% (Figure R5; Note that in the study we found that the median DNA methylation variance explained by a mQTL was 3.1%, with an interquartile range of 1.9%-6.3%).
The robustness of the East Asian specific mQTLs is also supported by the very high replication rate in another East Asian cohort (99.6% in CAS, FDR<0.05).

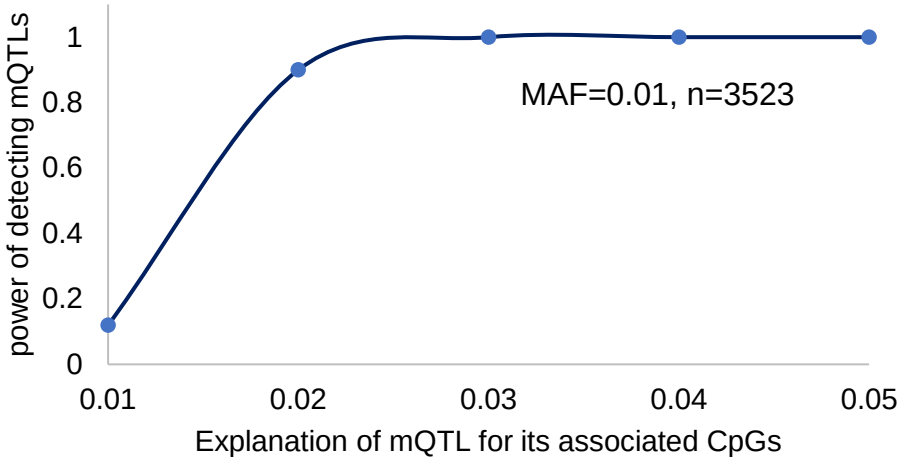


Figure R5 Power simulation of mQTL detection in NSPT (n=3,523) with MAF=0.01, variance explained 1% to 5%, P-value<=1×10^{-12}.

8.2 - The authors demonstrated that 81% of ~16M mQTLs with genome-wide significance in the FHS study [Huan et al. 2019] were also significant in the current East Asian cohorts at FDR < 0.05. It will be informative to understand what is the proportion in Huan et al. that achieved statistical significance for the genome-wide significant mQTLs in the current study, restricted to methylation markers on the 450K array.
Response: Thank you for pointing this out. We did restrict the comparison to methylation markers that overlapped between 850K and 450K arrays, which demonstrated that 81% of ~16M mQTLs with genome-wide significance in the FHS study[2].
In this revision, we also compared mQTLs detected in NSPT with those reported in the meta-analysis of European cohorts (GoDMC)[3] based on the overlapped SNP-CpG associations in both studies. At a genome-wide significance level (P-value < $1\lozenge10^{-14}$), we found 91% of mQTLs detected in NSPT were also replicated in GoDMC.
We have incorporated these new analyses in Results on Pages 8~9 'East-Asian specific mQTLs'.

8.3 - Tha authors focuses the study on the ethnic population (East Asians) that it was conducted in, as well as the potential role of mQTLs in mediating disease risk. It will be interesting to evaluate if the SNPs (and possibly CpGs) from the proposed panel of 'ethnic-specific' mQTLs are enriched for disease/phenotype traits that are known to differ in risk/prevalence between East Asians and Caucasians.
Response: Thank you for this valuable suggestion. We agree that more cross-ethnic analyses should be added, and we have done so in the revised manuscript. Please see our response to comment 1.

In particular, we found that the enrichment analysis of East Asian specific mQTLs in GWAS catalog implied a pronounced enrichment in diseases/traits with known prevalence differences between populations, e.g., attention deficit hyperactivity disorder characterized by a lower prevalence in East Asians than in Europeans[4], bipolar disorder with a lower prevalence in East Asians than in Europeans[5], pancreatic cancer with a longer survival in East Asians than in Europeans[6], and trans fatty acid levels with a substantially lower level in East Asians than in Europeans[7,8].

9. On a related noted, it will be informative to consider sex-specific mQTLs.
Response: The reviewer has raised a very interesting point. Since sex differences exist in most aspects of physiological and pathological processes, it would be helpful to demonstrate if there were sex-specific mQTLs and their potential contribution to the physiological and pathological differences between female and males. Here we preliminarily explored if there were sex-specific mQTLs in 60,490 independent trans-mQTLs. However, we found that there were only limited sex-specific mQTLs (0.01% female-specific and almost zero male-specific), implying that most mQTLs were common between female and male. Of course, this preliminary analysis has limitations as it focuses only on trans-mQTLs and the sex distribution of our samples is also biased (2,213 females and 1,310 males). In the future, we would consider validating sex-specific mQTLs based on a larger and balance-designed study.

10. The choice of samples for the East Asian validation dataset should be clarified. The samples were obtained from participants in two clinical trials of chiglitazar, which is tested for use in the treatment of Type 2 Diabetes. It is unclear from the current manuscript if blood was taken at baseline (before chiglitazar treatment) or after. However, it is most likely safe to assume that the participants are individuals suffering from T2D. We know that methylation levels can be modified by drug treatments, and also from previous studies that methylation profiles are altered in T2D patients, even before disease onset. The study will benefit greatly from a validation series that is population-based.
Response: We thank the reviewer for raising this point. DNAm associations with T2D that are not driven by changes in cell-type composition have not demonstrated the level of reproducibility seen for other phenotypes like aging, smoking or BMI, suggesting that DNAm changes associated with T2D, if any, will be a much lower effect size compared to the much

larger effect sizes associated with mQTLs. Indeed, the low level of reproducibility of reported T2D DNAm changes is a clear indication that effect sizes of true T2D-DNAm changes would be lower than 1%. This is because even 1% DNAm effect sizes (such as those associated with smoking) are highly reproducible. mQTL effect sizes are typically above 1-5% and many are even higher. In addition, it is worth pointing out that the samples from this cohort are all pre-treatment and all samples are from T2D patients, so there is no confounding by T2D/control status anyway. As with any other blood-based cohort, the major sources of variation in this dataset are (i) cell-type heterogeneity (variations in blood cell-type composition can be significant between any two individuals, healthy or not), and (ii) genetic (i.e. mQTLs). Thus, this cohort is entirely adequate for the purpose of validating mQTLs. Indeed, the very strong validation results obtained clearly support this. The reviewer's concern would be more justified had we seen a far from optimal validation.
Nevertheless, in response to the reviewer's point we have now performed the validation in an additional healthy East Asian cohort, a Han Chinese cohort from Beijing (CAS). It also indicated high replication rate for mQTLs detected in NSPT, 93.8% (FDR < 0.05), with high directional consistency (99.7%, r = 0.97, P-value < 1.00×10^{-323}). Compared with CGZ, CAS could replicate more mQTLs in NSPT (93.8% for CAS vs 87.1% for CGZ).
We added the new validation of CAS in Results on Pages 7 and Fig. 1e, and took the validation of CGZ into an Extended Figure (Extended Fig. 6).

11. There are also some concerns on the technical methodology. Firstly, the authors have opted for a 2-step analysis strategy for mQTL mapping, including the excluding of outlying methylation values in the 2nd step. It is unclear why the authors opted for this 2-step strategy, and also the reason for excluding 'extreme' methylation values (defined as outside the range of mean±3SD). This methylation values may well be the most informative, and the authors should also assessed the impact of this filtering.
Response: Thank you for pointing this out. We applied a two-step analysis strategy for mQTL mapping, including the excluding of outlying methylation values (outside the range of mean±3SD) in the 2nd step. While almost all of the mQTLs in two steps showed associations in the same direction (mQTLs in the same direction > 99.996%), only a small number of mQTLs that showed in different associations in two steps, as examples shown in Fig. S14. These paradoxical associations were possibly generated by DNA methylation stratification (Fig. S14). When excluding the outlying methylation values (outside the range of mean±3SD), the problem of DNA methylation stratification largely diminished (Fig. S14). The mQTLs calculated based on these exclusions reflect the effect of SNPs on the majority of DNA methylations in the blood. This strategy might be a little conserved but it could reduce false positives.
We added these results in sensitive analysis in Supplement.

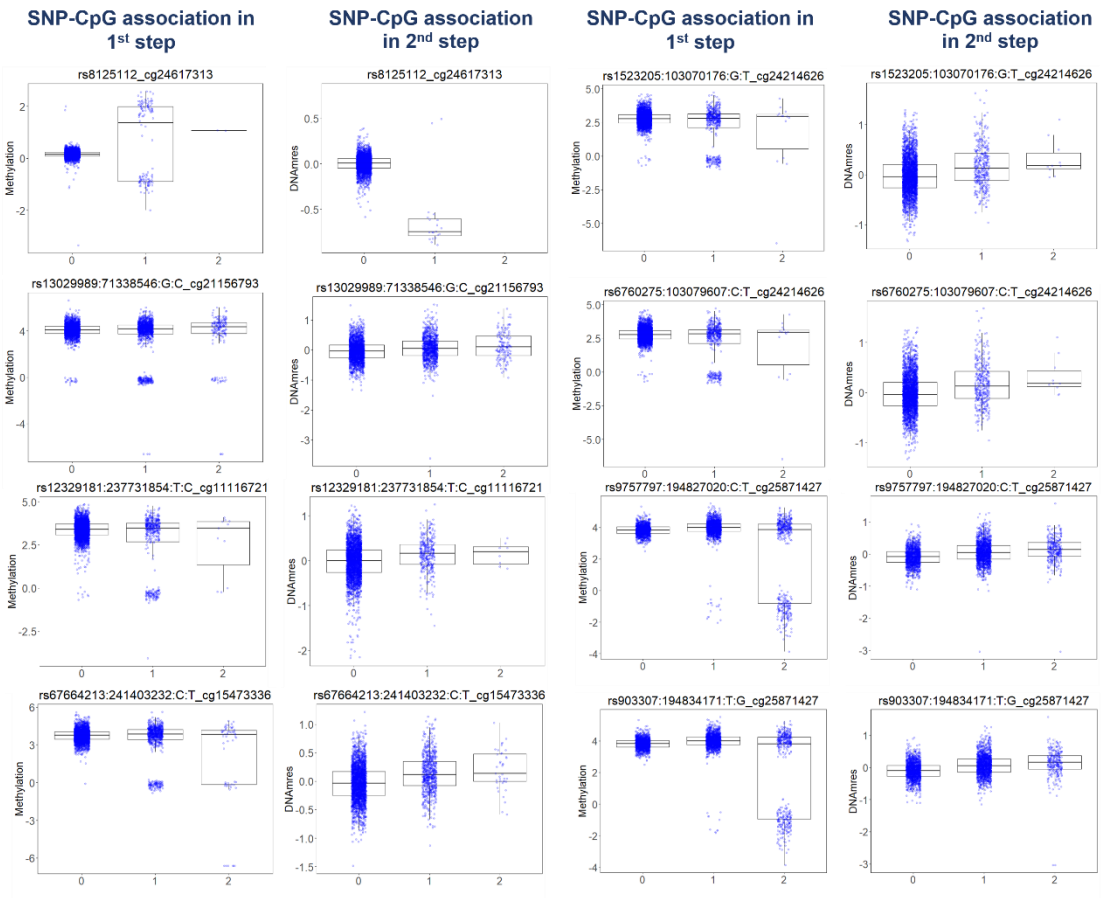


Fig. S14 SNP-CpG association shown in different directions for the same mQTL detected in 1st and 2nd step.

12. It is also stated that in the mQTL analysis, adjustments were made for 'bisulfite slide number' (please clarify what this refers to), batch (again, please clarify this), as well as the top 2 DNAm PCs. What is the rationale for adjusting for the top two DNAm PCs?

Response: Thank you for pointing this out. 'Bisulfite slide number' refers to the methylation array sentrix ID, and 'batch' refers to the batches of methylation experiments (Illumina EPIC), both of which are experimental factors. We considered to adjust DNAm PCs because we were concerned that confounding factors – measured or not – might interfere with the detection of mQTLs. Since the samples were collected from three regional districts of China, the environment and lifestyle in the local districts might be different. Nevertheless, we also generated mQTLs without adjusting two DNAm PCs, and found that the direction of these two sets of mQTLs were completely the same, except for slight differences in P values (Figure S15). This indicates that adjusting two DNAm PCs or not would not really affect the mQTLs discovered in our study. We have incorporated these sensitive analyses in Supplement.

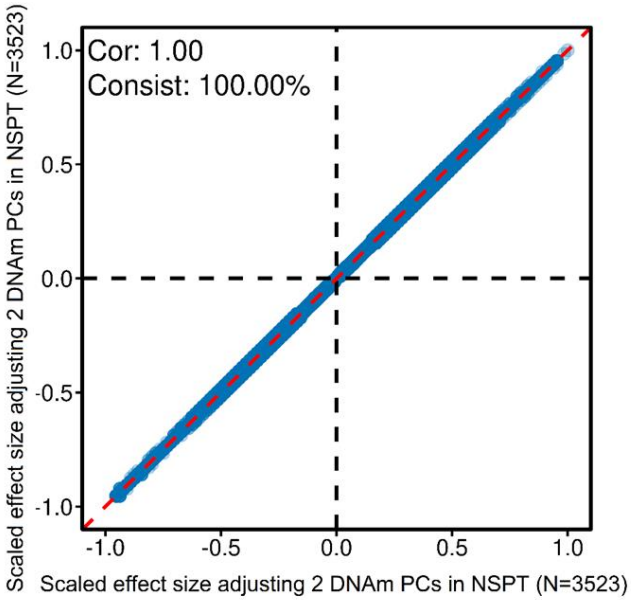


Fig S15 scatterplot of scaled effect size between mQTLs in NSPT (n=3,523) adjusting (x-axis) and not adjusting (y-axis) two DNAm PCs

13. In addition, it was stated in the Methods that missing beta values were imputed by impute.knn. Whilst it is not uncommon to impute for missing values, it is worth noting that in a recent comparison of methylation data imputation performances across seven methods [Lena et al. 2020], it was concluded that impute.knn is not suitable for DNA methylation data imputation, and that in general, it will be prudent to accompany data imputation by sensitivity analyses.
Response: The reviewer has raised a valid point and we thank the reviewer for drawing our attention to the paper by Lena et al. It is important to point out that the results obtained by Lena et al were obtained on studies profiling very small numbers of samples (typically on the order of 5 to 10 samples). In such a scenario, we would never advise running impute.knn to impute missing values, for the simple reason that impute.knn relies on computing correlations between probes across samples to reliably identify probes which can be used to impute values in a given probe. Such correlations would be extremely unreliable when computing them over only 5 to 10 samples. In our EWAS scenario we have over 3000 samples, and in this scenario our experience dictates that impute.knn works well. For example, we have carried out a simulation to assess the impact of sample size on the performance of impute.knn. As expected, this indicates that performance of impute.knn is much better in the large sample size setting (Figure S16). We have added this figure to the Supplement.

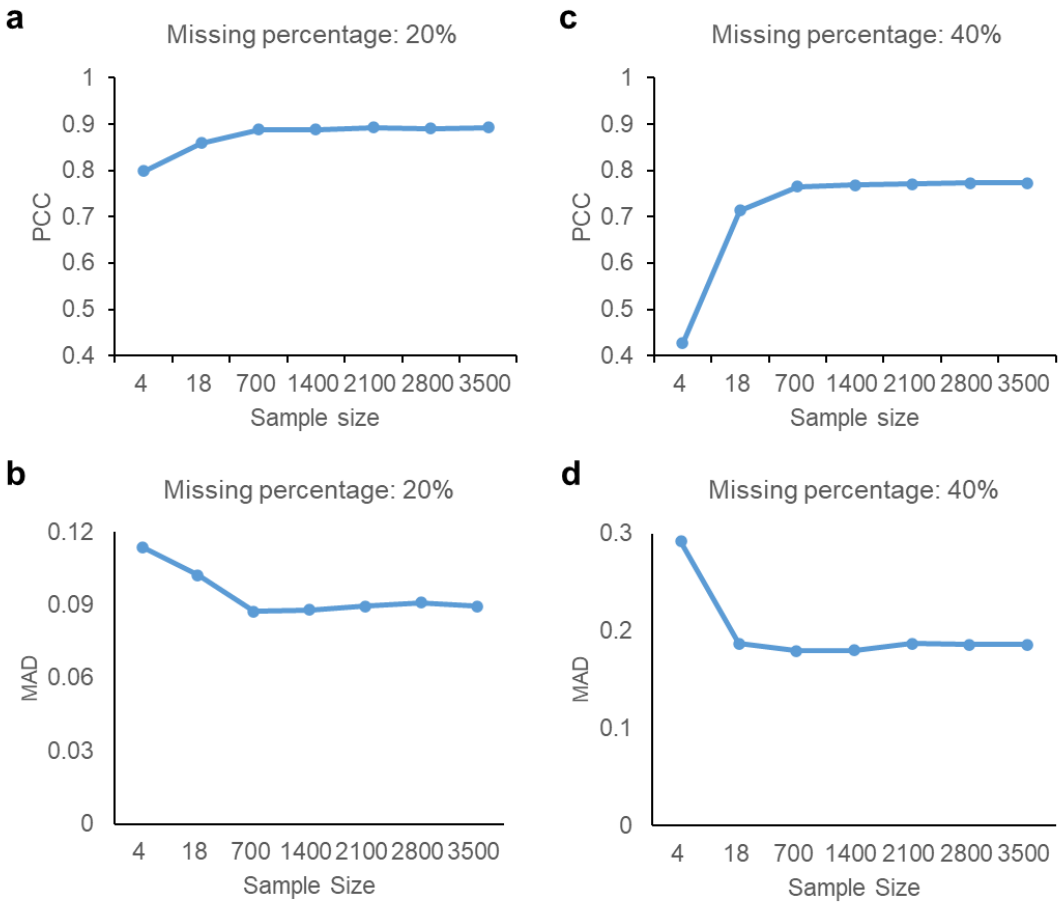


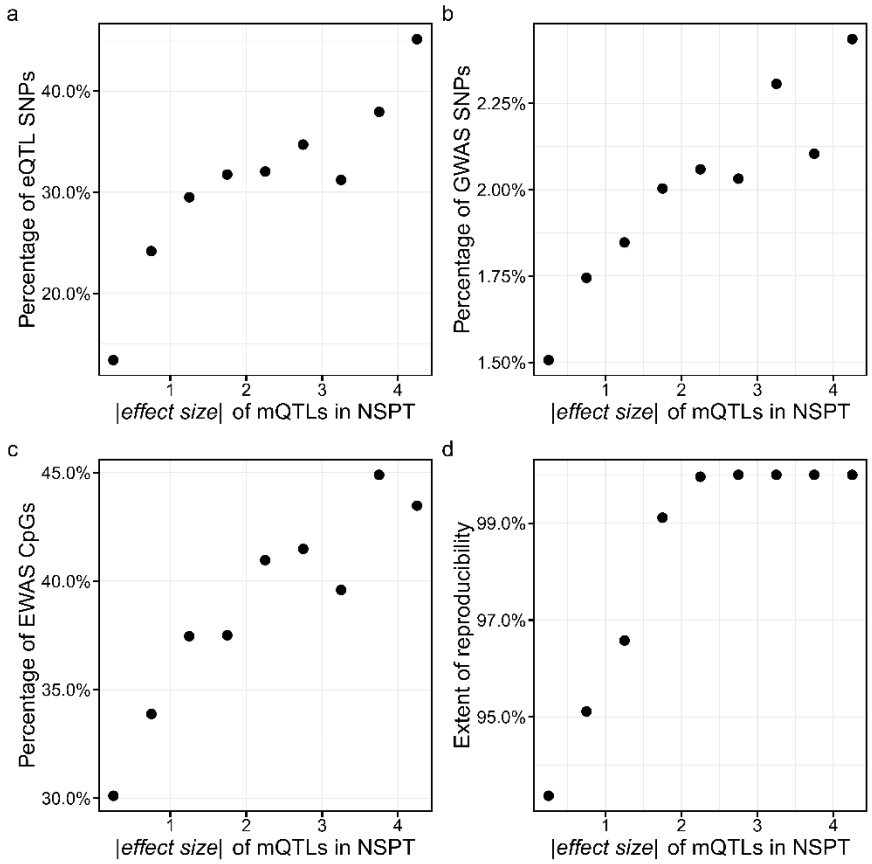
Figure S16 Performance (Pearson correlation and MAD between imputed values and original values) of impute.knn in simulation scenario (missing rate 20%, 40%, 100 randomly selected CpGs)

From Lena et al, we also learned that whilst impute.knn displays larger RMSE for extreme DNAm values (near 0 or 1), it displays the best RMSE at moderate DNAm values. In our data, mQTLs are more likely to be associated with CpGs that display moderate DNAm-levels, hence impute.knn appears to be a good imputation method for the goal of detecting mQTLs.

14. The selection of control/background group is critical in enrichment analyses. For instance, for the enrichment analysis of mQTLs in 3D chromatin contacts, the authors defined the control groups as i) random sampling of SNP-CpG pairs from all SNP-CpG combinations, regarded as genomic background and ii) SNP-CpG pairs with the same distance distribution as mQTL pairs (distance-matched SNP-CpG). This represents rather loose matching criteria, without taking into consideration other important factors such as MAF of SNP and variation of methylation level at

CpG.
Response: Thank you for this advice. We have constructed a strict background to conduct the enrichment analysis. Specifically, to take the MAF of SNP and the variation of CpG methylation level into consideration, we only involve SNPs and CpGs that appears in mQTLs to generate control groups.
The results are updated in Results on Pages 14~15 and Figure 3a.

15. For the audience to better appreciate the mQTLs, it will be helpful for the authors to quantify the methylation effect size, and also with respect to whether the effect sizes were associated with stronger biological implication (e.g. association with gene expression and disease/phenotypic traits), as well as extent of reproducibility.
Response: Thank you for these valuable suggestions. Following these suggestions, we quantified the methylation effect size. We found that the median absolute change in methylation M value per allele copy was 0.12 (interquartile range 0.08-0.21). We described the relation of effect size between mQTLs, eQTL, GWAS signals and mQTL replication. In most cases, as the effect size of mQTL increased, it was more likely to be eQTLs (**Extended Fig. 4**a), GWAS SNPs (**Extended Fig. 4**b) and with high reproducibility **Extended Fig. 4**d), and we also discovered that the associated mCpGs were more likely to be EWAS signal (**Extended Fig. 4**c).
We have incorporated these analyses in Results on Page 5~6 in '**mQTL mapping, annotation and replication**' and **Extended Fig. 4**.



Extended Fig. 4 The relation between effect size of mQTLs and their biolgicial implications and reproducibility.

a, the proportion of mQTLs which are also cis-eQTLs. **b,** the proportion of mQTLs which are also GWAS signals. **c,** the proportion of mQTLs associated CpGs which are EWAS signals. **d,** the proportion of mQTLs which are replicated in CAS (n=1,060) with the same direction and FDR<0.05.

16. Finally, I believe that the study will be greatly strengthened by wet-lab experimental validation that could provide convincing evidence of the proposed causal role that mQTLs play, and/or the central role of transcription factors in bridging genetic variants and methylation levels.
Response: Thank you for this suggestion. We agree that wet-lab experimental validations could strengthen some of the findings. However, due to the limitation in budget and time, we are unable to carry out wet-lab experimental validations in this study.

Reviewer #2:
Remarks to the Author:
1. The authors are undertaking a worthwhile examination of the genetic variants that control

DNA methylation in an East Asian population, to address concerns in the field over differences in this regulation by genetic race as well as differential genetic regulation by cell subtype. They demonstrate extensive overlap of mQTLs with a prior study in a European ancestry population, and using innovative cell specificity estimation demonstrate substantial overlap of these traits across cell types. They go on to describe the potential influence of chromatin architecture and develop a better understanding of how trans-mQTLs may be operating, and link their findings to two important phenotypes. These are significant and original findings with implications for downstream research efforts.
Response: Thank you for the appreciation of our work.

2. Title: Suggest adding that this is "in blood", eg. "Comprehensive mechanistic characterization of mQTLs in blood in an Asian population, " given the cell type specific nature identified. It is likely that there may be further differences in other tissues that were not examined in this study.
Response: We appreciate the reviewer's point. We agree that adding the specific tissue would be helpful. However, since the majority of mQTL studies have been based on blood samples, and "blood" has rarely appeared in the title, we think it might not be necessary to emphasize on it in our title as well. Therefore, we'd prefer to keep the title 'Comprehensive mechanistic characterization of mQTLs in an East Asian population'. We did emphasize that the mQTLs are in blood in various places in the manuscript.

3. Methods: The authors have applied state of the art methodologies in the determination of mQTLs, including developing a novel C++ based algorithm to enhance the speed of detection of those mQTLs and reduce computing burden. The application of the CellDMC method to estimate cell type specificity of the mQTLs is also highly innovative, and provides important additional information regarding cellular specificity of this genetic regulation. All statistical methods and inference appear appropriate and robust, and the incorporation of a number of large, publicly available datasets adds to the value of these findings and their interpretation.
Response: Thank you for the appreciation of our work again.

4. The examination of FOSL1 and NFKB1 hotspots add to the understanding of trans-mQTLs and disease process. It would be helpful, though, to have a better understanding of why these two hotspots, of the 16 top 1% hotspots were chosen for further dissection. Were the other hotspots not in disease associated regions?
Response: Thank you for pointing this out. We chose these two hotspots as FOSL2 at H2 was the most significant from TFmotifView, and NFKB1 at H5 was the most significant from PWMEnrich. Trans-mCpGs were significantly enriched in the binding sites of the TFs near their trans-mQTLs, which are often known susceptibility loci of human traits and diseases.
There are other hotspots overlapped with known trait-related regions like hematological measurement, cardiovascular disease, inflammatory measurement and so on (Table R1). Especially, we found that trans-mQTL hotspots were enriched with hematological traits related variants. Together with another discovery of our work, that trans-mQTL hotspots enriched with

super enhancers, we proposed that trans-mQTL hotspots play key roles in maintaining cell identity. And the story between FOSL2-mediated trans-mQTL hotspot and eosinophil counts provide a complete chain of evidence for this functional link. We have updated the related description in Results on Pages 21~22 and added a supplement table (Table S5) to demonstrate the relation between trans-mQTLs hotspots and diseases/traits.

Table S5. The 16 hotspots index mQTLs or their high LD (r^2>0.8) SNPs in GWAS Catalog.

			GWAS Catalog hits		
HotSpot	**index mQTL**	**N LD mQTL**	**Hematological**	**Inflammatory/immune**	**Cardiovascular/metabolic**
H4	rs58408429	25	Y	Y	Y
H5	rs3774937	16	Y	Y	Y
H6	rs651297	14	Y	N	N
H8	rs143396005	31	N	N	Y
H9	rs1038353	26	Y	Y	N
H10	rs10883359	19	Y	Y	Y
H11	rs3809627	11	Y	Y	N
H12	rs17818238	16	Y	N	N
H13	rs11082385	40	Y	Y	N
H15	rs28789846	23	Y	Y	Y
H16	rs7275212	13	Y	Y	N

N LD mQTL: the number of mQTLs that in LD r^2>0.8 with index mQTL.

5. Discussion: In the examination of mQTLs and hotspots, the authors performed a mendelian randomization and concluded that the methylation at these regions does appear to be in the causal path of the outcomes examined (eosinophilia, obesity). This is in contrast to prior work on mQTLs (Min et al, 2021) which concluded that in most cases methylation was not causally mediating a variety of traits, including blood specific traits. The authors should discuss their findings in light of these results.

Response: Thank you for these suggestions. We agree that in most cases methylation was not causally mediating a variety of traits. In line with this, for the two examples in our study, we did only find a small proportion of the trans-mCpGs that might be the potential causal factors to eosinophilia or obesity. The conclusions in our study were not in contradiction with that in Min et al (2021). We noticed that the results from Mendelian Randomization tests (MR) might be biased due to weak instrument, confounding factors or pleiotropy between exposures and outcomes. The conclusions from MR need to be further validated. However, the latest mQTL paper in Hawe et al (2022) did report that trans-mQTLs associated CpGs were likely to be causes of BMI[10], which partly supported our discoveries. Although our causal BMI mQTLs did not overlap with the causal BMI mQTLs identified by Hawe et al (2022), our strategy to focus on trans-mQTLs co-localizing with TF-binding could represent a novel paradigm for identifying

potentially causal mQTLs. In support of this, we note that several of the identified mCpGs map to genes (e.g., NOD2, PTPN3) that in mouse models have already been causally implicated in obesity[11-14]. We note that these findings do not contradict the dearth of casual associations reported recently by Min et al (2021)[1], as this latter study did not specifically focus on trans-mCpGs co-localizing with TF-binding.
We have updated these in Discussion on Page 29.

Reviewer #3:
Remarks to the Author:

A. Summary of the key results

1. The authors describe the largest mQTL-mapping study in a Han Chinese population (n=3523) using DNA methylation measured in whole blood. The find over 80% of mQTLs in common with a similarly-sized white population (FHS, n=4170) and replicate 87% in smaller Han Chinese population (n=798). They apply CellDMC to their whole blood data to identify cell-type specific mQTLs and estimate that <10% of mQTLs are cell-type specific. They confirm the importance of transcription factors to the functional roles of trans-mQTLs and explore roles for DNA methylation in mediating the effects of trans-mQTL 'hot spots' on eosinophilia, ulcerative colitis and body mass index.
Response: Thank you for the appreciation of our work.

B. Originality and significance: if not novel, please include reference

1. This is the first significant mQTL-mapping study in an Asian population. To the reviewers knowledge, the largest previous Asian mQTL study included Chinese (n = 93), Indians (n = 83) and Malays (n = 78):

Kassam, I., et al (2021). Genome-wide identification of cis DNA methylation quantitative trait loci in three Southeast Asian Populations. Human molecular genetics, 30(7), 603–618.
Response: Thank you for the appreciation of our work.

2. Although cell-type specific mQTLs are reported, these were estimated from whole blood DNA methylation using the CellDMC software tool. Methods like CellDMC are still relatively new and untested. Preliminary evaluations and the validation reported in this manuscript indicate that any reported cell-type specific associations should be considered highly speculative. These cell-specific results should not therefore be considered a significant contribution to the literature.
Response: We appreciate the reviewer's point, but the argument for not attempting to identify cell-type specific mQTLs is in our opinion very weak. Ideally, we would perform the mQTL

analysis in cell-sorted samples for 3 or more blood cell subtypes across 3000 samples, but doing so is logistically impossible and very costly. Contrary to what the reviewer is stating, algorithms like CellDMC have been extensively tested [15,16]. Although we agree that the sensitivity to detect cell-type specific mQTLs in less frequent or less variable cell-types is very limited, these previous studies have shown that CellDMC can reliably detect cell-type specific DNAm signals at the resolution of 2 to 4 cell-types/lineages. As a concrete example, in buccal swabs (which consist of 50% immune-cells) CellDMC correctly predicted smoking-associated DNAm changes in the immune-cell compartment which have been consistently found to be associated with smoking in blood. Given that the effect sizes of mQTLs are typically around 10 to 20% (over 97% of our mQTLs display a DNAm change per allele copy less than 0.1 or less than 0.2 if measured between the A/A & B/B genotypes), i.e. much larger than the effect sizes associated with smoking, it makes a lot of sense to attempt finding cell-type specific mQTLs. That we have been able to identify and validate a number of cell-type specific mQTLs in blood is in our opinion worth reporting. In order to alleviate the reviewer's concern we have added a new panel Fig.3f to provide clear examples of CellDMC mQTL predictions, and glancing at the DNAm profiles of these mQTLs fully confirms that CellDMC is working well. For convenience we display this panel below:

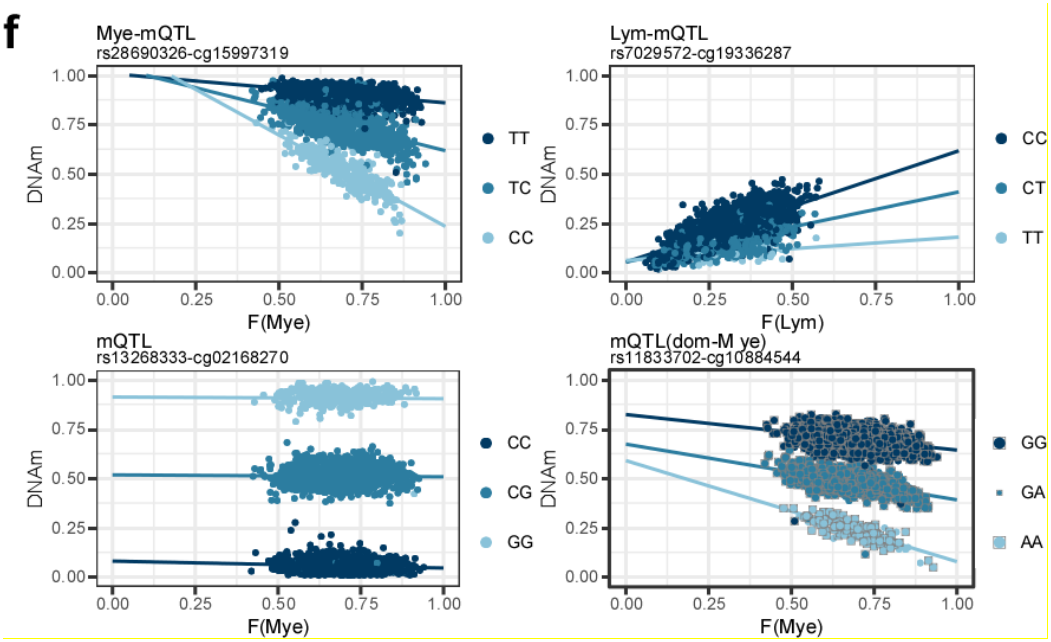


Figure 3f Scatterplots of DNAm (y-axis) vs cell-type fraction F (x-axis) for 4 mQTLs with samples colored by genotype.

The top two mQTLs are examples of a myeloid and lymphoid-specific mQTL, with the x-axis labeling the myeloid and lymphoid fraction, respectively. The bottom two mQTLs are examples of two cell-lineage independent mQTLs, with the left mQTL being equally dominant in myeloid and lymphoid subsets and the right mQTL being more dominant in the myeloid subset.

However, we take the reviewer's point on board, and in response we have clarified in Discussion that our cell-type specific mQTL findings need to be interpreted with caution and that they await further validation.

3. The manuscript concludes with the claim that the described mQTL database is "an invaluable resource for understanding the genetic and epigenetic variations in disease predisposition between ethnic groups." Although this is likely true, the analyses of the manuscript mainly focus on mQTLs in common with previous European studies (e.g. trans-mQTL hotspot relevance to disease). I was expecting the study to focus on Asian-specific mQTLs and their potential role in diseases with higher prevalence in Asian populations.
The authors note that a variety of factors other than ethnicity could explain differences between their study and previous non-Asian studies (Line 414 "e.g., differences in power or Illumina beadarray version"). Although this is true and lack of a significant p-value should not be used conclude absence of an association, it still possible to compare mQTL effects between studies and note where effects are significantly different.
Response: Thank you for this valuable suggestion. We agree that East Asian specific mQTLs and their potential roles in diseases with different prevalence in populations should be considered. We performed a cross-ethnic comparison using the recent large-scale meta-analysis of European cohorts from the Genetics of DNA Methylation Consortium (GoDMC)[3]. These are the main findings: 1) Among the 2.65 million NSPT mQTLs, the majority of them (2.41 million or 91%) were also study-wide significant mQTLs in GoDMC. The fact that majority of the mQTLs are not population specific holds true regardless of the significance threshold used. 2) The remaining 9% (238K) mQTLs, regarded as East Asian specific mQTLs in NSPT, could be replicated very well in another East Asian cohort CAS (99.6% replicated). 3) For a mQTL in East Asians, the likelihood of it being also a study-wide significant mQTL in Europeans heavily depended on its MAF in the Europeans, and vice versa. 4) The enrichment analysis of East Asian specific mQTLs in GWAS catalog implied a pronounced enrichment in diseases/traits with known prevalence differences between populations (e.g., attention deficit hyperactivity disorder[4], bipolar disorder[5], pancreatic cancer[6], and trans fatty acid levels[7,8]). These significant findings support that the presentation of population specific mQTLs might due to natural selection on allele frequencies in different populations and further contributes to the prevalence/risk difference of some common complex diseases among populations.

We have incorporated these new analyses in Results on Pages 8~9 '**East-Asian specific mQTLs**' and also modified our Discussion on Page 27~28.

C. Data & methodology: validity of approach, quality of data, quality of presentation

1. The mQTL analysis appears to have been sound.

Response: Thank you for the appreciation of our work.

2. The cell-count specific mQTL analyses are highly speculative because, as noted earlier, the methods are still new and relatively untested and performance evaluations indicate high error rates. The manuscript should be more clear in showing how, although there is some evidence of validation, the validation is extremely limited and shows much higher error rates than we'd expect if the analyses has been performed in purified cell-type populations.

Response: We appreciate the reviewer's point. Although we have already addressed a similar comment earlier (see our response to reviewer's comment B2), here we would like to add the following. The validation of the cell-type specific mQTLs in the independent EA-cohort (see Fig.3c-d) is in our opinion remarkably strong, specially when the inference is made at the resolution of two cell-lineages (myeloid vs lymphoid). In the BLUEPRINT cell-sorted data we are also able to validate cell-type specific mQTLs (Fig.3e), despite the fact that we did not have BLUEPRINT summary statistics for all our SNP-CpG pairs and despite the much smaller sample sizes in BLUEPRINT (n~200). Thus, whilst the validation is not as strong as when validating mQTLs in tissue, they are significant and hence worth reporting.

In addition, when the reviewer talks about "high error rates" it is extremely important for the reviewer to clarify if he/she is referring to sensitivity, specificity or precision, since an algorithm like CellDMC exhibits different levels of sensitivity, specificity and precision, depending also on the particular constellation of cell-type specific DNAm changes that can arise. It is worth highlighting here that the complexity of identifying cell-type specific DNAm changes is huge: sensitivity, specificity and precision is a function of (i) sample size, (ii) number of cell-types in the tissue, (iii) the number of cell-types exhibiting a DNAm change, (iv) their effect sizes, (v) their direction of effect (hypermethylated or hypomethylated), (vi) the mean level and variance of cell-type fractions, specially for those cell-types that are altered, (vii) the presence of additional confounders and (viii) whether the effect is homogeneous within all cells of a given cell-type. Given that the biological ground truth for cell-type specific mQTLs is largely unknown, it is extremely difficult to make predictions of what the expected sensitivity, specificity and precision would be. In Zheng et al (2018)[15], we do provide extensive simulation analyses, which support the view that the algorithm is sufficiently powered and that it displays high specificity for scenarios where DNAm changes happen in one or more cell-types. In You et al (2020)[16] we also presented additional simulations, demonstrating that in the simple case of 2 major cell-types or lineages (e.g. myeloid vs lymphoid), that detecting a cell-lineage independent DMC separately in the two lineages requires lots of samples, specially for the cell-type/lineage that is least abundant and which displays lower variance. Based on all these extensive simulations, we can summarize CellDMC's performance as of (i) reasonably high sensitivity to detect cell-type specific DNAm changes in frequent or moderately frequent cell-types, (ii) low sensitivity to detect cell-type specific effects in minor or less variable cell subpopulations, and (iii) high specificity and precision, which contradicts the claim of high type-1 error rate implicit in the reviewer's criticism. In other words, when CellDMC discovers an mQTL that is present in CD4+ T-cells but not in other cell-types, we can be fairly certain that the mQTL is indeed present

in CD4+ T-cells (as clearly shown by our validation in BLUEPRINT FACS-sorted data, as well as in the independent EA-cohort), but we are less certain about the negative findings in the other cell-types (e.g. CD8+ T-cell, B-cells…), because the power or sensitivity to detect them in the other cell-types could be quite low. Once again, we take the view that the results and the positive validations obtained in BLUEPRINT and in the independent EA-cohort are worth reporting, but agree with the reviewer that we need exercise caution when interpreting the findings.
Below, we provide a number of examples of myeloid-specific, lymphoid-specific, common and predominantly myeloid or lymphoid mQTLs to help the reviewer better appreciate that the CellDMC predictions are well supported by the patterns of DNAm change (Figure R6-R8). We have incorporated these figures in Supplement.

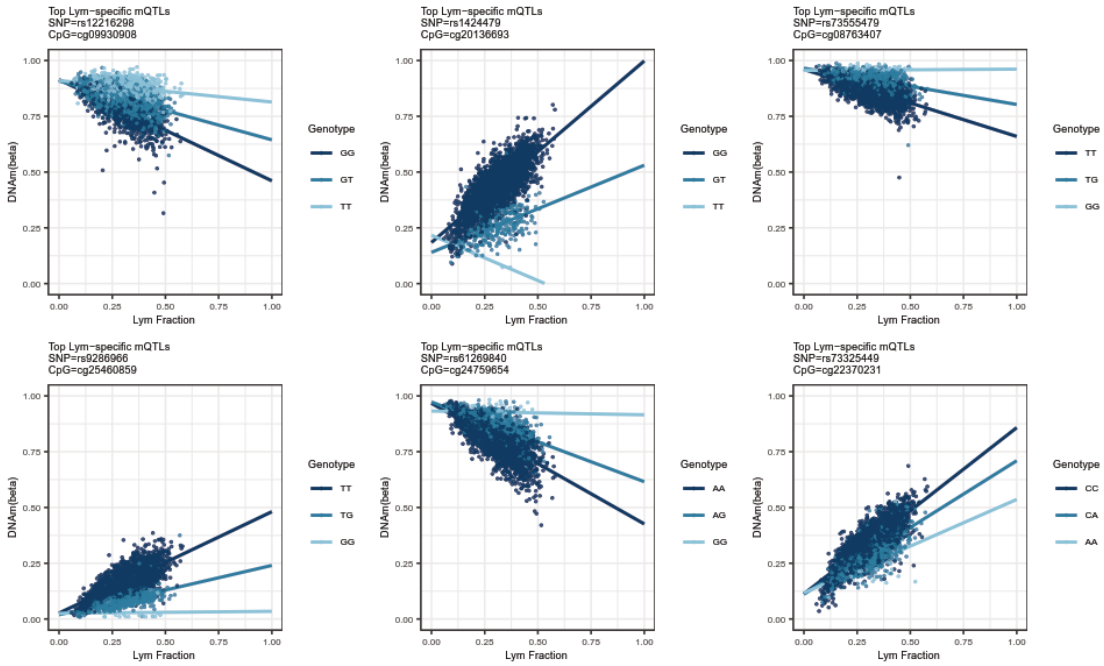


Figure R6 Examples of significant lymphoid lineage specific mQTLs.
The x-axis shows the lymphoid lineage fraction of samples and the y-axis shows the DNA methylation level of the mQTLs associated CpGs.

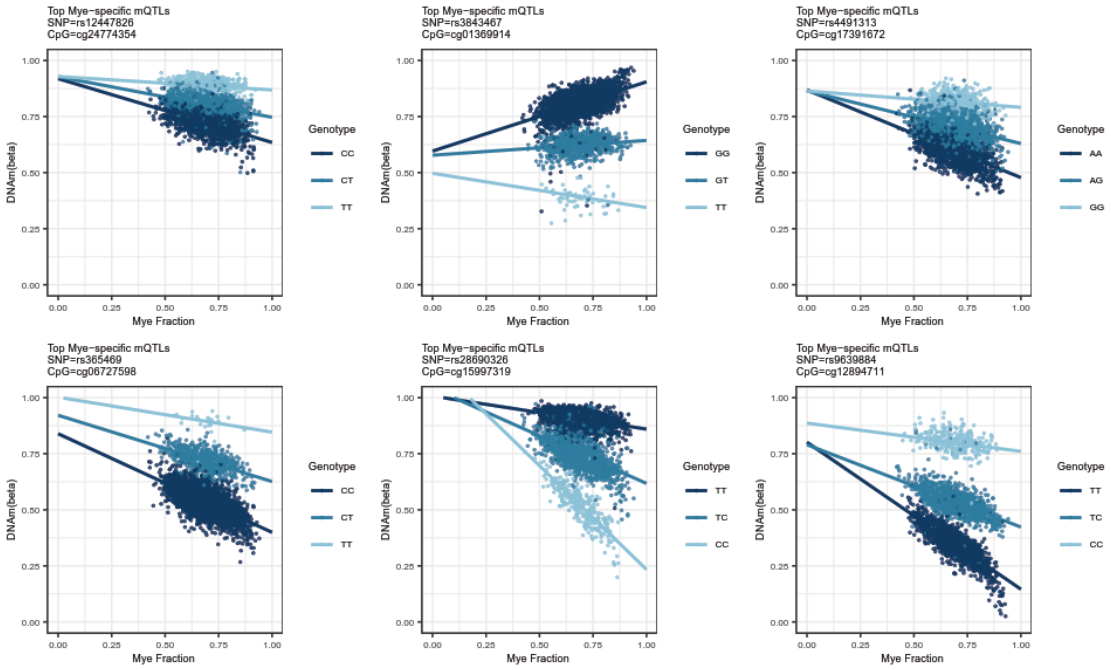


Figure R7 Examples of significant myeloid lineage specific mQTLs.
The x-axis shows the myeloid lineage fraction of samples and the y-axis shows the DNA methylation level of the mQTLs associated CpGs.

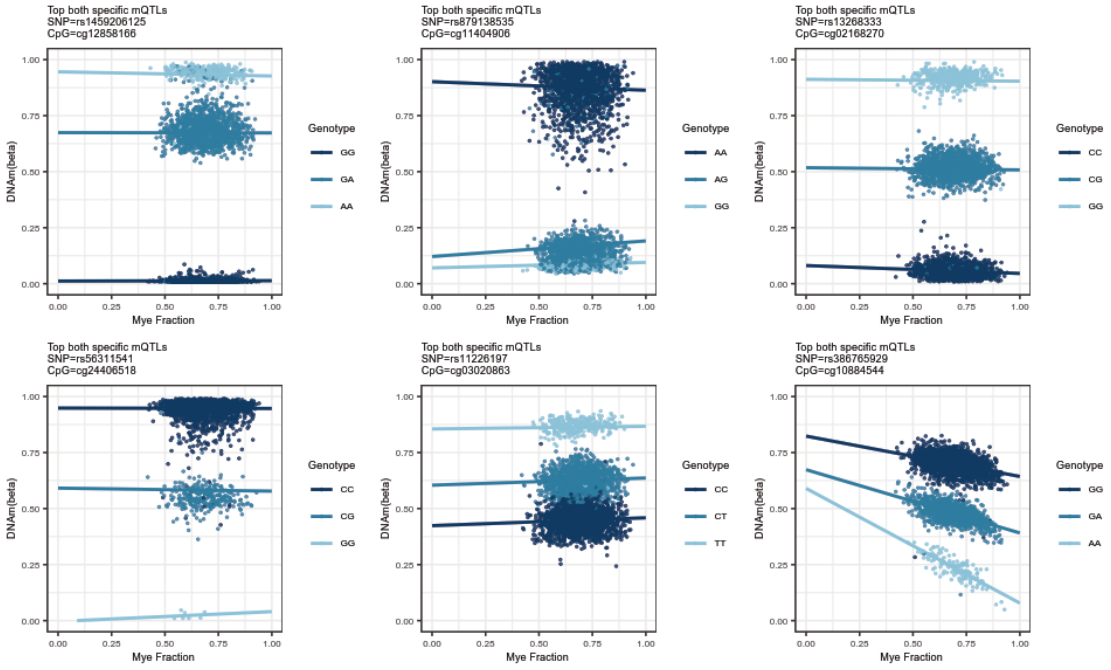


Figure R8 Examples of significant non-specific mQTLs.

The x-axis shows the myeloid lineage fraction of samples and the y-axis shows the DNA methylation level of the mQTLs associated CpGs.

3. It should be expected that more cell-type specific associations should be observed in the more abundant cell types as cell-type specific signal in the bulk tissue data will be stronger. For some reason the authors use this observation to conclude that the more abundant cell types are more 'dominant in blood' (lines 165-168 and lines 171-173). Besides the uncertain meaning of 'dominance' in this case, this limitation of the data should not be used to draw biological/functional conclusions.
Response: We think that the reviewer has misunderstood the meaning of our statements on lines 165-168. These statements as well as the accompanying data displayed in old Fig.2f (this data has now been placed in Extended Fig.8) represent a form of validation: if we detect a mQTL in the myeloid-lineage (which makes up 69% of blood), then on average it should display a higher effect size than a mQTL we detect in the lymphoid lineage, for the simple reason that myeloid cells make up a much higher fraction of cells in blood. Thus, old Fig.2f (new Extended Fig.8) is meant to demonstrate that the CellDMC predictions of what are myeloid and lymphoid-specific mQTLs is consistent with their expected effect sizes. In response to the reviewer's point, and because we agree that this data constitutes more of a verification or confirmation of the predicted cell-type specificity, we have decided to place this old Fig.2f in the Supplement.

4. The analysis of chromatin accessibility requires rationale for what seem to be arbitrary decisions:
4.1 - Are models 1 and 2 the only possible models? How were they selected? The criteria for each appears to be quite specific.
Response: Thank you for raising up the questions. We acknowledge that the two models we proposed are not the only possible models. Since there were no available models for us to select, we proposed these two models based on the characteristics of mQTLs we saw in this study and our understanding about the mechanisms of genetic regulatory processes from SNP to CpG.

4.2 - What proportion of the the mQTL associations should we expect to explain based on models M1 and M2. The analyses suggest that models M1 and M2 "explain 40% of mQTLs". Is this more than expected? Should we be proposing additional models to explain the remaining 60%?
Response: Thank you for these valuable suggestions. To check the proportion of mQTLs expected to be explained, we generated a control SNP-CpG group by sampling SNP-CpG pairs with the same distance distribution as mQTL pairs. And to take the MAF of SNP and the variation of CpG methylation level into consideration, we only involve SNPs and CpGs that appears in mQTLs to generate the control group. Then, we calculated the number of SNP-CpG pairs that quantified the constraints of the M1 or M2 model, as we did for mQTL pairs in the main text. The result shows that 19.9% of the SNP-CpG pairs in the control group can be explained by the proposed models, that is, the proportion of explained mQTL pairs (40.4%) was

2.03-fold than that of the control group. Finally, we acknowledge that there are still undiscovered mechanisms between mQTLs and their mCpGs which can further be proposed.
We modified the corresponding description in Results on Pages 14~15 and also in Discussion on Page 28.

5. The "OpenCausal tool" is applied with little explanation or rationale. The text should include a short introduction to what the tool is, how it assesses causal relationships and the limitations of those assessments.
Response: Thank you for pointing this out. Following the reviewer's advice, we added the rationale of OpenCausal in Results on Page 15, 'Relying to OpenCausal database[17] which predicted the change of chromatin accessibility scores based on TF expression and SNP background before and after SNP mutation to measure the influence of a variant on the regulatory element (see Supplement), we found that our cis/lcis-mQTLs were significantly enriched in blood-specific opening-causal SNPs (2.0-fold, Fisher $P=1.6\times10^{-16}$; Fig. 4f).'.
In this study, we collected the processed fragments per kilobase million (FPKMs) of RNA-seq data for blood tissue from GTEx project. Using the TF expression and mQTL information as inputs, OpenCausal calculated an opening-causal score for each mQTL. We consider the mQTLs with non-zero scores as the ones that are sensitive to the chromatin accessibility. One limitation of this assessment is that it only focuses on the influence of the given SNP on the regulatory region, without considering the possible combining impact of nearby SNPs. But this assessment still suits our situation, as we only want to evaluate the influence of each one mQTL on the chromatin of its local region.
Detailed description of OpenCausal was included in Supplement.

6. The manuscript claims to provide evidence for "DNAm levels at NFKB1 trans-mQTLs being causal mediators for BMI, as opposed to being a consequence of BMI" (Lines 371-372). First, I think the statement should refer to DNAm levels at mCpGs of the NFKB1 trans-mQTLs. Secondly, and more importantly, this finding appears to contradict an extensive literature on DNA methylation in blood and BMI, including the Wahl et al and Mendelson et al (Plos Med 2017) studies, which find almost no evidence for a causal effect on BMI. More generally, Min et al (2021) report, based on a much larger sample size (n=30K), very little evidence for a causal effect of DNA methylation on any phenotype. The authors should more carefully investigate these appearent disagreements with previous studies. It isn't sufficient to just note that Min et al "did not specifically focus on trans-mCpGs co-localizing with TF-binding." The Min et al study was genome-wide and better powered, so it should have identified at least as many causal relationships.
Response: Thank you for these suggestions. We agree that in most cases methylation was not causally mediating a variety of traits. In line with this, for the two examples in our study, we also found that there were only a small proportion of the trans-mCpGs might be potential causal factors to eosinophilia or obesity. The conclusions in our study were not in contradiction with that in Min et al (2021). We noticed that the results from Mendelian Randomization tests (MR)

might be biased due to weak instrument, confounding factors or pleiotropy between exposures and outcomes. The conclusions from MR need to be further validated. The latest paper in Hawe et al (2022) also reported that trans-mQTLs associated CpGs were likely to be causes of BMI[3], which partly supported our discoveries. Although our causal BMI mQTLs did not overlap with the causal BMI mQTLs recently identified by Hawe et al (2022), our strategy to focus on trans-mQTLs co-localizing with TF-binding could represent a novel paradigm for identifying potentially causal mQTLs. In support of this, we note that several of the identified mCpGs map to genes (e.g., NOD2, PTPN3) that in mouse models have already been causally implicated in obesity[11-14]. We note that these findings do not contradict the dearth of casual associations reported recently by Min et al (2021)[1], as this latter study did not specifically focus on trans-mCpGs co-localizing with TF-binding.
We have updated the Discussion on Page 29.

7. A prominant claim in the paper is that clusters of trans-mQTLs tend to coincide with clusters of transcription factor binding sites. In the text, the authors confusingly refer to this as trans-mQTLs being "surrounded by TFs". The supplementary methods defines this as being located within 1Mbp of a predicted transcription factor binding site. If this indeed the definition, then the text should just clearly state this simple definition rather than leave it buried in the supplementary materials. I'm not sure that this definition makes any sense. How likely is it that a genetic variant will influence the binding of a transcription factor 1Mbp away? I would have expected a distance with a much smaller distance.
Response: Thank you for pointing this out. We added the definition in the text. We also tested the results at smaller distances, i.e., 1Kb, 10Kb and 100Kb, and all the results were similar to the original result with 1Mb. We chose 1Mb to show the results because this is the commonly used cis-region definition which can cover more potential associations[18-22].
We have added this in Results as Fig. 5d.

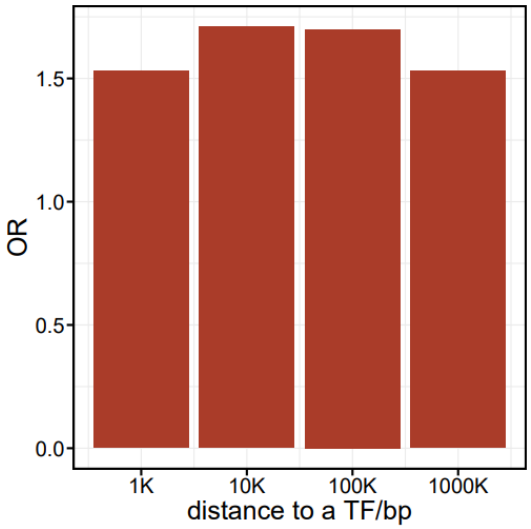


Fig. 5d The relative ratio of trans-mQTLs located within 1Kbp, 10Kbp, 100Kbp and 1Mbp of a predicted transcription factor coding region.
We used the Fisher's exact test to check if it's more likely to find a TF near a trans-mQTL than near a random SNP from the background SNPs in NSPT. Odds ratios (OR) in 4 different distances (i.e., <1K, 10K, 100K and 1Mbp) to a TF.

8. One piece of the evidence supporting this claim is the scatterplot in Figure 4b that is claimed to show a correlation between the number of trans-mQTLs and the number of transcription factor binding sites on the same chromosome. The correlation between the two appears to be driven mainly by chr19. What is the correlation with and without chr19?
Response: Thank you for pointing this out. We are sorry for the misleading statements. Without chr19, there was indeed no correlation between TF density and number of trans-mQTLs on chromosomes (r=0). We have deleted these results in the revised manuscript. Instead, we focus on TF density and trans-mQTL hotspots as hotspots have always been the focus in the study. We found that there was a positive correlation between hotspot density and TF density on each chromosome, even after excluding chr19 ($P=1.23\times10^{-4}$ and r=0.71 for all chromosomes, $P=5.52\times10^{-4}$ and r=0.66 for excluding chr19).
We have updated our description in Results on Page 18.

D. Appropriate use of statistics and treatment of uncertainties

1. The authors report 62.92M mQTLs. It is standard to also report the number of independent mQTLs.
Response: Thank you for pointing this out. We found 56.29M cis-, 2.27M lcis- and 4.36M trans-mQTLs in whole genome. After pruning redundant SNPs for each mCpG in each category by limiting LD to r2<0.2, there remained 1.75M independent cis-, 52.25K lcis- and 111.63K trans-

mQTLs.
We added this description in Results on Page 5.

2. In many places p-values well below the precision of floating point calculations are reported (e.g. p < 1.00×10^-323). Values this small are meaningless and should be replaced with a value that better represents the capabilities of the computers used for analysis (e.g. a typical recommendation is p < 2.22x10^-16). Statistical strength of associations with extremely low p-values is better expressed with summary statistics such as effect sizes and confidence intervals.
Response: Thank you for pointing this out. We appreciate the reviewer's point, yet this is a very subjective point as many journals, including top-ranking ones allow the quotation of P-values even to levels as low as 1×10^{-300}. Whilst we agree that P-values can never be computed to this precision level, it is nevertheless true that a P-value estimated to be 1×10^{-100} is more significant than one estimated to be 1×10^{-15}, for the simple reason that the statistic associated with 1×10^{-100} has a larger absolute value compared to the statistic associated with a P-value of 1×10^{-15}. In fact, the two-tailed P-value is in 1-1 correspondence with the absolute value of the statistic, not the effect size. We can have mQTLs with very low effect sizes but very high statistic values (and very low P-values) if the variances are really small, and conversely a large effect size may not necessarily be very significant if the variances within genotype are large. To resolve the reviewer's point, one could display the actual statistic values, but for the general readership this will be harder to interpret. For instance, nobody would know what the P-value of a t-statistic that takes the value 10 is. Most people know that an absolute t-or z- statistic value close to 1.96 corresponds to a P-value of 0.05, but hardly anyone remembers what the P-value of t=10 is. For this reason, and although there is an error-bar associated with quoting P-values that are highly significant, this is still preferable and more informative than vaguely stating $P<2\times10^{-16}$, because if a P-value is 1×10^{-200}, the associated absolute statistic is definitely much larger than the statistic associated with say a P-value of 1×10^{-16}. Hence, based on this very rationale, we'd prefer to keep the quotation style of P-values.

3. Many enrichment analyses are reported, in most cases using the hypergeometric test. Authors should take care to correctly specify the universe/background in these tests. In most enrichment tests, the text does not clearly indicate how the universe/background was defined. For example, line 124-125 says that "mSNPs were enriched in genomic functional regions such as promoters and exons, and this pattern was more pronounced for trans-mSNPs than cis-mSNPs". It is unclear whether or not this enrichment accounted for the fact that DNA methylation measurements on the Illumina Beadchips are highly enriched in promoters and exons.
Response: Thank you for this suggestion. We have modified our statements about the background used in enrichment analyses in the manuscript. The enrichment analyses were accounted for the proportion of background SNPs or CpGs in related functional regions. The fold changes reflected the enrichment of cis-, lcis, trans-mQTLs, or mCpGs compared with background SNPs or CpGs in each functional region.

E. Conclusions: robustness, validity, reliability

1. Conclusions about cell-type specificity should be strongly qualified in light of the methods used and validation findings.
Response: Thank you for this suggestion. We added discussion about the methods and results of cell-specificity in the context.

F. Suggested improvements: experiments, data for possible revision

1. Overall, the manuscript text needs to be revised to use technical terms correctly and precisely and to simplify text that is unnecessarily complex.
Response: Thank you for this suggestion. We have revised our manuscript carefully, corrected technical terms and simplified the text.

2. The chromatin analysis needs to be better explained and justified.
Response: Thank you for this suggestion. We followed the reviewer's advice and added more explanation and justification in the chromatin analysis. We added background SNP-CpG groups to show what proportion of mQTLs is expected to be explained, and carried out enrichment analysis with more strict criteria of SNPs and CpGs.

3. To capitalise in the major contribution of this study, mQTLs in an Asian population, the authors should make an effort to identify ethnicity-specific mQTLs and investigate their potential role in ethnicity-specific disease.
Response: Thank you for this suggestion. We agree that more cross-ethnic analyses should be added, and we have done so in the revised manuscript.
We detected 238K East Asian specific mQTLs through a cross-ethnic comparison between East Asian (NSPT) and European (GoDMC). We also explored the enrichment of the East Asian specific mQTLs in GWAS catalog which implied a pronounced enrichment in diseases/traits with known prevalence differences between populations, e.g., attention deficit hyperactivity disorder characterized by a lower prevalence in East Asians than in Europeans[4], bipolar disorder with a lower prevalence in East Asians than in Europeans[5], pancreatic cancer with a longer survival in East Asians than in Europeans[6], and trans fatty acid levels with a substantially lower level in East Asians than in Europeans[7,8].
We have incorporated these new analyses in Results on Pages 8~9 '**East-Asian specific mQTLs**' and also modified our Discussion on Page 27~28.

G. References: appropriate credit to previous work?

1. Shoud cite the largest previous Asian mQTL study included Chinese (n = 93), Indians (n = 83) and Malays (n = 78) and compare findings:

Kassam, I., et al (2021). Genome-wide identification of cis DNA methylation quantitative trait loci in three Southeast Asian Populations. Human molecular genetics, 30(7), 603–618.
NG; 2022
Response: Thank you for pointing this out. We had cited this work in Introduction on Page 4.

2. It is somewhat unexpected that findings are not compared to the largest mQTL study carried out so far (n=30K, Min et al. 2021).
Response: Thank you for pointing this out. Following this suggestion, we compared our mQTLs with this recent meta-analysis of European cohorts (GoDMC).
We have incorporated these new analyses in Results on Pages 8~9 '**East-Asian specific mQTLs**' and in Discussion on Page 27~28.

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

1. A lot of the text in the results section is unnecessarily complex. For example, consider the following sentence on lines 100-101:
"The mQTL SNPs (mSNPs) covered more than 2/3 of tested SNPs (5.56M), while mQTL CpGs (mCpGs) covered 1/3 of tested CpGs (284,128)."
Here is a simpler version:
"Two-thirds of the tested SNPs (5.56M) were associated with DNA methylation, while one-third of tested CpG sites (284,128) were associated with genetic variation."
This example highlights two causes of unnecessary complexity that appear repeatedly throughout the results section. The first cause is the misuse of terms already well-defined in the literature. In this example, term "mQTL SNP" is redundant because an mQTL is by definition a SNP, a SNP that is associated with DNA methylation at a CpG site. Thus, "mSNP" is an unnecessary definition because mQTL and mSNP are equivalent. The second problem is unusual choices of words and phrases. The term "covered" here is confusing because it suggests a more complex relationship between mQTLs and SNPs than that mQTLs are simply a specific subset of SNPs that are associated with DNA methylation. The results section needs to be revised to simplify the text and ensure correct use of defined terms.
Response: Thank you very much for these helpful suggestions. We have revised the manuscript carefully and made changes according to these suggestions.

2. Another important example is references to trans-mQTL "hot-spots" which are very simply defined as genomic loci containing a large number of trans-mQTLs. However, confusion is caused by reference to a "trans-mQTL network" (e.g. Line 255) which is never defined and to

"unlinked trans-mCpGs" whose vague definition is buried in the supplementary materials. There CpG sites are mysteriously "clumped" in 500Kbp windows to "exclude linkage among adjacent CpGs". The term "linked" here is non-standard and actually incorrect because it refers to "linkage disequillibrium" (LD). LD is about genetic variation, not DNA methylation variation. It is more typical to refer to an "index" CpG site which representats a cluster of strongly correlated CpG sites. The "hotness index" would then be defined for a cluster of trans-mQTLs in linkage disequillibrium as the number of associations of these mQTLs with trans index CpG sites.
Response: Thank you very much for these helpful suggestions. We have revised the manuscript carefully and rewritten these parts according to these suggestions.

Below are other examples:

3. Line 92 algorithm, fastQTLmapping was up to 4 and 11 times faster in the single-thread and 32 CPU threads
"up to" isn't very meaningful, summarize with ranges or averages
Response: Thank you for pointing this out. We agree that it should be changed to “ranged from 4 to 11 times”. However, this part has been removed from the revised manuscript.

4. Line 119 allele frequency differences for pan-ethnic mSNPs were significantly smaller when compared to the

I assume the allele frequency differences referred to are between FHS and EAS.
Response: Yes, it is the allele frequency differences referred to between FHS and EAS.

5. Line 120 18.91% mSNPs that were only significant in FHS

This is an unnecessary use of jargon. Better to say "mSNPs that were only observed in FHS" something like that.
Response: Thank you. We have modified this sentence.

6. Line 194 cis-mQTL pairs

I suspect that this refers to pairs of associated CpG sites and cis-mQTLs. However, this term is not correct.
Response: Yes, it refers to cis-mQTLs and their associated CpGs. We have modified the description in Results.

7. Line 256 With hotness-index
Line 257 increasing, the proportion of mSNPs in hotspots surrounded by TFs was monotonically increasing

Simpler to write: As the hotness-index increases, the proportion of trans-mQTLs within 1Mbp of a transcription factor binding site increases monotonically.
Response: Thank you. We have changed the description according to this suggestion.

8. Figure 2a doesn't add up. The myeloid/lymphocyte analysis reports about 2E7 mQTLs in lymphocytes but about twice that number in specific lymphocyte cell types. By contrast 6E7 mQTLs are reported in myeloid cells but about half that number in specific myeloid cell types.
Response: Thank you for pointing this out. The mQTLs displayed in each column was obtained by interaction analysis between SNPs and corresponding cell components independently. There was not additive relationship between the upper and the lower panels because the mQTLs discovered in these specific cells might be overlapped at varied degrees.

9. Line 310 The majority (141, 60.8%) of the 232
Line 311 mQTLs were detected exclusively in the myeloid lineage (Fig. 5b&c, Fig. S31b).

Okay, but how strong is the evidence that they do not occur in lymphocytes? There appears to be something wrong with Figure 5b. There are 232 mQTLs, 141 are detected in myeloid cells but none in lymphocytes?
Response: Yes, there are 141 mQTLs detected in myeloid cells but none in lymphocytes according to CellDMC at a strict significance threshold. When filtering cell lineage mQTLs at the nominal threshold (0.05), there are 230 left in myeloid cells and 7 in lymphocytes. For cell types with small proportion in blood, we thought it was not powerful enough to discover more cell-type specific associations in our study.
We updated the Figure and the description in Results on Page 21~22, where Figure 5b was taken out, and the description relevant to cell-lineage specific was also deleted.

10. Line 312 Atlas31, we found the 232 mCpGs to be mainly enriched in immune system disorders (P.bfadjust =

I assume that "P.bfadjust" is just refers to a Bonferroni adjusted p-value.
Response: Yes, it is.

11. In Figure 5f, "Kim data" should be replaced with a more formal citation of the dataset.
Response: Thank you, we have changed it into a formal citation.

12. Line 356 Another important trans-mQTL hotspot was driven by a GWAS SNP associated with ulcerative colitis
Line 357 (UC) and linked in-cis with the transcription factor NFKB1

What does it mean for a hotspot to be 'driven by' a specific SNP?
Response: Thank you for pointing this out. To make a clear statement, we modified the

description about this part in Results on Page 24, '*NFKB1* at H5 on chr4 represents our most significant finding from PWMEnrich (**Table 1**). The index mQTL rs3774937 at H5 was significantly associated with ulcerative colitis (UC) in GWAS Catalog ($P=5.0\times10^{-8}$).'

13. Line 358 al (2017)1, and thus validating this NFKB1 trans-mQTL network in an EAS population

Unsure where "trans-mQTL network" is defined.
Response: Thank you for pointing this out. We have changed it into 'trans-mQTL hotspot'.

14. Line 364 that a list of 364 CpGs known to be associated with BMI (as derived from Wahl et al (2019)64 and Line 365 other studies), did so also in our Asian cohort and with the same directionality of DNAm change

Simpler to say "that published associations of 364 CpG sites with BMI (Wahl et al 2019) were replicated in our Asian cohort"
Response: Thank you for pointing this out. We have modified the sentence according to your suggestion.

15. Line 419 associated with environmental factors.

I don't understand this sentence.
Response: Through population-specific mQTLs analysis, we found that a large proportion of mQTLs are pan-ethnic, and the mechanisms underlying the formation of mQTLs unlikely differs between ethnic groups. But for disease predisposition, mQTLs is not enough, it should be combined with ethnic and individual exposome to unravel the pathogenesis of disease.
We deleted this sentence in this revision, to make our discussion more compact.

16. Line 792 g, Enrichment of mQTL pairs in functional elements.
Line 794 Heatmap shows the fold changes (see Methods) of SNP-CpG pairs in all combinations of functional
Line 795 categories.

Fold changes with respect to what?
Response: Here fold changes refer to the proportion of the mQTLs and their associated CpGs mapped to functional categories compared with that of randomly selected SNP-CpG pairs. We compared mQTLs and their associated CpGs belonging to each functional category with randomly selected SNP-CpG pairs, and we found that mQTLs and their associated CpGs were more likely to both map to transcription factor binding sites (TFBS), to promoters, and to CTCF binding sites.
We have updated these descriptions in Results on Page 5.

References:

1. Min, J.L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet* **53**, 1311-1321 (2021).
2. Huan, T. *et al.* Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat Commun* **10**, 4267 (2019).
3. Hawe, J.A.-O.X. *et al.* Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat Genet* **54**, 18-29 (2022).
4. Fayyad, J. *et al.* The descriptive epidemiology of DSM-IV Adult ADHD in the World Health Organization World Mental Health Surveys. *Atten Defic Hyperact Disord* **9**, 47-65 (2017).
5. Zhang, L. *et al.* The prevalence of bipolar disorder in China: A meta-analysis. *J Affect Disord* **207**, 413-421 (2017).
6. Rawla, P., Sunkara, T. & Gaduputi, V. Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World J Oncol* **10**, 10-27 (2019).
7. Jiang, L. *et al.* Trans fatty acid intake among Chinese population: a longitudinal study from 1991 to 2011. *Lipids Health Dis* **19**, 80 (2020).
8. Restrepo, B.J. Further Decline of Trans Fatty Acids Levels Among US Adults Between 1999-2000 and 2009-2010. *Am J Public Health* **107**, 156-158 (2017).
9. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
10. Hawe, J.S. *et al.* Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat Genet* **54**, 18-29 (2022).
11. Gurzov, E.N., Stanley, W.J., Brodnicki, T.C. & Thomas, H.E. Protein tyrosine phosphatases: molecular switches in metabolism and diabetes. *Trends Endocrinol Metab* **26**, 30-9 (2015).
12. Rodriguez-Nunez, I. *et al.* Nod2 and Nod2-regulated microbiota protect BALB/c mice from diet-induced obesity and metabolic dysfunction. *Sci Rep* **7**, 548 (2017).
13. Gurses, S.A. *et al.* Nod2 protects mice from inflammation and obesity-dependent liver cancer. *Sci Rep* **10**, 20519 (2020).
14. Kreuter, R., Wankell, M., Ahlenstiel, G. & Hebbard, L. The role of obesity in inflammatory bowel disease. *Biochim Biophys Acta Mol Basis Dis* **1865**, 63-72 (2019).
15. Zheng, S.C., Breeze, C.E., Beck, S. & Teschendorff, A.E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods* **15**, 1059-1066 (2018).
16. You, C. *et al.* A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. *Nat Commun* **11**, 4779 (2020).
17. Li, W., Duren, Z., Jiang, R. & Wong, W.H. A method for scoring the cell type-specific impacts of noncoding variants in personal genomes. *Proc Natl Acad Sci U S A* **117**, 21364-21372 (2020).
18. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
19. Díez-Villanueva, A. *et al.* Identifying causal models between genetically regulated methylation patterns and gene expression in healthy colon tissue. *Clin Epigenetics* **13**, 162 (2021).
20. Nalls, M.A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* **46**, 989-93 (2014).
21. Liu, N.Q. *et al.* The non-coding variant rs1800734 enhances DCLK3 expression through long-range interaction and promotes colorectal cancer progression. *Nat Commun* **8**, 14418 (2017).
22. Xu, J. *et al.* Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biol* **20**, 243 (2019).

Decision Letter, first revision:

9th Aug 2022

Dear Sijia,

Your Article, "Comprehensive mechanistic characterization of mQTLs in an East Asian population" has now been seen by the original 3 referees. You will see from their comments below that while they find your work has improved in revision, some important points are raised. We remain interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

In brief, Reviewers #2 and #3 - who were previously supportive and critical, respectively - are now satisfied and supportive of publication.

However, Reviewer #1 - while appreciating the revision - thinks that their major comments have been not been comprehensively addressed. They think there is a lack of striking, EAS ancestry-specific novelty despite this being a major strength of your dataset; that the claims for mQTL cell type specificity need to be further supported by additional analysis; and a lack of overall biological novelty, in the context of those last two points.

We think that Reviewer #1's comments are reasonable and we do agree that these aspects could be improved. We think that Reviewer #1 provides specific guidance that, to our reading, would not necessarily require a substantial further expansion of the work (e.g. additional novel data generation). It's less clear to us whether all three of these aspects need to be substantially improved or whether a focus on one may convince this reviewer, but we would highly recommend the focus on the EAS ancestry aspects of your work, as this is what makes your manuscript most distinctive in comparison to the current literature.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

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

Important:
You have also been in touch stating that you'd like this manuscript to be published as soon as possible. Given that Reviewer #1 still has novelty concerns that we are, at this stage, not prepared to overrule, I am also consulting with my colleagues at Nature Communications to see whether they would offer an Accept in Principle decision without requiring a further major revision (as we would

need). I have not yet heard from them, but I will forward their feedback once I have.

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

When revising your manuscript:

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

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

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

Please be aware of our guidelines on digital image standards.

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

[Redacted]

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

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

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

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

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

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
I thank the authors for the revised manuscript, which and the accompanying rebuttal of reviewer concerns.
The manuscript and materials are extensive, as is often the case for these complex mQTL studies. I note that the study is better couched now in terms of global literature, in particular with analysis of complementary data from Europeans (although not South Asians). I do have some persisting significant concerns:
1. Importance of the study for East Asian populations.
The majority (91%) of East Asian mQTLs are also found in Europeans. The 9% that show less evidence for replication have low AF in Europeans. mQTLs in blood thus appear to be generalizable across populations, as previously demonstrated (Hawe et al, 2022). My question about the significance of the East Asian specific mQTLs remains just as valid, but has not been answered well. The only test for the role of these East Asian specific mQTLs appears to be an enrichment test in the GWAS catalogue. Based on this the authors suggest 13 GWAS traits (out of how many? May hundreds presumably?) are enriched at just $P<0.05$. This does not appear to be Bonferroni corrected, or even FDR. Based on the y axis for Figure 2c, it might even be a one-sided test, to maximise permissivity. Since the GWAS summary stats are biased by European results and Afs, how was this addressed? Based on the P value distribution, it is unclear that any of these are notable statistically. Even for these mQTLs there is no evidence of the authors drilling down further into what genetic variants that seem to have greater relevance for East Asians. If these are Asian specific functional variants … what are they, and what are they doing? I still cannot see how this study improves my understanding of health or biology in East Asians.

2. Cell-specific effects.
The issue about cell-type specific effects is very interesting, and can reveal new biological insights, including into disease pathways. The authors do not seem to have taken on board though the criticism of only looking for cell-specific effects in the mQTLs from whole blood. The authors take the view that this is a 'very minor limitation'. In doing so, they propose that most mQTL are cell-type independent. As one line of evidence of this they use the Hawe et al study which looks at cosmopolitan mQTLs from across populations and tests them in white cell subsets and isolated adipocytes. I would highlight that this test of generalizability in Hawe et al was limited to mQTLs selected to be shared. Without

acknowledging this limitation, the authors are presenting what is a circular argument: things discovered as shared appear to be shared. BLUEPRINT is underpowered for this analysis. The authors propose that 90% of mQTLs are shared between cell subsets, and make the impressive claim: 'The key question is whether we are identifying the ~10% of cell-type specific mQTLs, and the answer here is yes'. The data presented do not support this view. Without a search for cell-specific mQTLs outside of the 'whole-blood set', the authors cannot know how many they are missing. The authors argue that it is computationally intractable. Well, at least do some random sets of apparently unrelated SNP-CpG pairs, to get a better estimate of the issue.

3. New biology
P13 – the role of TADs and shared chromatin state is reported.
P17 – the role of TFs as mediators of mQTLs is very well reported already.
P22-25 – provides 2 pathway analyses, to suggest a role for FOSL2 (eosinophils) and NFKB1 (BMI) as transacting mQTLs in respective traits. 2 stage MR is done, but not SMR or co-localisation analyses to assess whether the same genetic mechanism underlies the methylation and phenotypic traits. I think these kind of analyses are pretty standard now and should have been included. There are no experimental studies to validate the statistical observations. Since the East Asian and cell-specific aspects of the manuscript are not compelling, the potential value of the manuscript currently rests on whether the biology is a rigorous, interesting and important advance.

Reviewer #2:
Remarks to the Author:
The authors have done a remarkable job in addressing each of the criticisms raised by the reviewers, and the additions they have provided demonstrate the validity and quality of their approach, the robustness of their analyses and results, and clarify statements made and jargon.

Reviewer #3:
Remarks to the Author:
The authors have thoroughly addressed all of my comments.

I would just add that it looks amateurish to report p-values as small as p=1e-200 or p=1e-300. There is no meaningful difference between p-values this small due to how close they are to the limits of machine precision and the fact that underlying analyses rarely satisfy the assumptions of statistical tests to the precision implied by these p-values. I'm not aware that the journal has a policy about reporting small p-values, but they really should address this.

Author Rebuttal, first revision:

[insert PDF from NG-A58885R2]

Decision Letter, second revision:

10th Jan 2023

Dear Sijia,

Happy New Year! I would like to apologise to you and your co-authors for the prolonged review process, and thank you for your patience.

Your Article, "Comprehensive mechanistic characterization of mQTLs in an East Asian population" has now been seen by the original Reviewer #1. You will see from their comments below that while they appreciate the improvement in this revision, there are yet important comments that remain to be satisfactorily addressed. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

In brief, Reviewer #1 sounds - as yet - unconvinced of the overall novelty and utility of your study's findings. They think that the biological insights gained from the cis-mQTL analysis are limited, but suggest that analysing mQTLs based on their cis-/trans- mode of action and identifying trans-acting pathways may improve this. They remain unconvinced of the cell-specific analysis, but also sound as if there a way to improve this analysis also. They also suggest that experimental follow-up of the FOSL2 finding should be added.

Given the late stage of review and the support from the other two reviewers, we would like to avoid another major, time-consuming revision, if at all possible. However, we also think that Reviewer #1 does make some very useful suggestions, especially those comments on examining mQTLs split by their mode of action, and the cell specific analysis (which, in our reading, still requires a lot of improvement including further computational work). We are willing to overrule some of their requests, e.g. the experimental work, but we would like you improve the computational analyses as directed.

We appreciate that you and your co-authors would like to publish this study promptly. If you think that these requests will take too long and would prefer a speedier publication, please do get in touch - I would be happy to consult with my colleagues at Nature Communications regarding that.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to identifying key priorities that should be addressed in revision and sometimes overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritized set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

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

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

When revising your manuscript:

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

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

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

Please be aware of our guidelines on digital image standards.

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

[redacted]

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

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

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

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

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
I thank the authors for their detailed responses to my previous comments. The manuscript and responses are extensive, and represent a substantial volume of work.

Picking up on some of the results from the new analyses presented:

1. EA specific mQTLs. These are a small fraction of the total mQTLs at functionally relevant SNPs identified by GWAS (2.6% for the GWAS catalog, and 3.0% for BBJ GWAS data). Since 97% of the mQTLs that appear to be functionally relevant are already discovered / shared with other populations, the EA data could be seen as an incremental change in knowledge rather than fundamental new insight.

2. cis vs trans mQTLs. The bulk of the reviewer response talks about mQTL relationships without making the cis- trans- distinction. These are fundamentally different in terms of biological mechanism and interpretation. Trans-CpGs provide the opportunity to link remote loci into co-ordinate pathways. Many would argue that this opportunity to unravel trans-acting nuclear pathways has been the primary highlight from mQTL analyses. By comparison, the insights from a cis-mQTL are more modest (unclear even, see point 3). This limits interpretation of the material; for example when discussing cell-specific mQTLs, is the evaluation of cis-mQTLs, trans-mQTLs or a mix (if so, what mix?). Similar for the relationships of the 3% of mQTLs that are EA 'specific' – cis, trans, and any inference from this?

3. Integrating mQTLs with GWAS. The authors report 98 loci at which a SNP associated with phenotype shows colocalization with mQTLs in EA but not EUR. On this occasion, the text does make clear the analysis is of cis-mQTLs. However, the results show that the 'ethnic specific effect' is simply determined by differences in allele frequency (the SNP is low frequency or absent in Europeans; Fig 2). How does the presence of a cis-mQTL finding advance understanding at these loci? Taking the ELF1 locus as an example, the association of the locus with height is demonstrated by the SNP (and supported by rare genetic variants) in EA, and the role of the gene is demonstrated by ELF as an eQTL. How does the presence of some co-localising cis-mQTLs enhance understanding of the ELF locus? What is the genomic mechanism or process revealed? Co-localisation may indicate a shared underlying genetic basis, but does not indicate whether methylation is cause, consequence or covarying with the phenotype. More importantly, given the low coverage of the EPIC array (~3% of CpGs), the analyses do not identify the specific 'functional' CpGs or the 'functional' SNP, or the potential genomic mechanism linking SNP to methylation. This extends the point that, while trans-mQTLs can reveal regulatory pathways, the presence of a cis-mQTL is currently less informative.

4. Cell specific effects (1). The issue about shared vs cell specific effects remains problematic. Hawe et al identify a set of mQTLs in whole blood, and show that many replicate in adipose tissue. It remains the case that the use of a mixed tissue such as whole blood for discovery will favour identification of

mQTLs with low heterogeneity of effect between different cell types (ie that are shared). Respectfully, the reviewer was not confusing 'shared between ancestries' with 'shared between cell types'.

5. Cell specific effects (2). I was initially unclear how the authors were defining a cell specific effect. The text states that 'many mQTLs significant in one-lineage also displayed associations in the other, albeit marginally so'. Their approach seems to be to define a cell-specifc mQTL as i. being present in one cell type at P<10-8, but ii. no evidence for association (failing to reach P<0.05) in the alternate cell subset(s) [is this any or all subsets?]. Based on these criteria, they conclude that mQTLs are shared across tissues. It is helpful that the approach has been clarified. The result aligns with population genetic studies (for example) which show that the great majority of genetic associations in one ancestral group can also be demonstrated in other ancestry at a permissive threshold of P<0.05. However, this approach fails to address the critical biological question – which of the mQTLs play a more important biological role in one cell subset that another? For example, is the functional consequence of genetic variant in NFKB1 the same in all cell subsets or is it greater in immune cell subsets, and if so which one? Such an analysis would provide real new insight into cell specific biology, analogous to the insights generated by cell specific studies of gene expression (eg GTEX). Instead, the present analysis adopts an approach that blurs the distinction between cell subsets, and creates the impression that mQTLs are the same across cell-subsets.

6. New biology (1). Both shared CpG states at TADs and TF hotspots are recognized. In particular the 'hotspots' (eg compare Fig 5a with results in Bonder et al, Hawe et al). This is being oversold.

7. New biology (2). The trans-pathway linked to FOSL2 and blood counts is interesting, and would be compelling if supported by experimental evaluation of the statistical inferences. NFKB1 and BMI is less clear. If the NKFB1 SNP that drives the methylation in trans is not associated with BMI (line 1-2, page 28), then it is difficult to understand how the data suggest that NFKB1 trans-CpGs are causal in obesity

Author Rebuttal, first revision:

Point-to-point response to Reviewer-1 comments:

General Comment: I thank the authors for their detailed responses to my previous comments. The manuscript and responses are extensive, and represent a substantial volume of work.
Response: We thank the reviewer for engaging with our manuscript and for recognizing the substantial revisions made in response to his/her comments.

Comment-1. EA specific mQTLs. These are a small fraction of the total mQTLs at functionally relevant SNPs identified by GWAS (2.6% for the GWAS catalog, and 3.0% for BBJ GWAS data). Since 97% of the mQTLs that appear to be functionally relevant are already discovered / shared with other populations, the EA data could be seen as an incremental change in knowledge rather than fundamental new insight.
Response: We appreciate the reviewer's comments on EA-specific mQTLs, but respectfully disagree with the interpretation that these findings only represent an incremental change in knowledge.

1. First of all, we would like to emphasize the importance of human diversity in epigenome studies as recently highlighted by Breeze et al (2022)[1]. Indeed, our study contributes the first large mQTL mapping in Han Chinese, which is a significant contribution in itself. Without such a study, we would not know whether the majority of mQTLs are shared across different ethnic populations. Our demonstration that most mQTLs are shared between ethnic groups provides valuable information for future studies, enabling researchers to focus their efforts on combined mQTL-eQTL mapping and facilitating the setup of international consortia to improve the identification of ethnic/ancestry-specific mQTLs.
2. Whilst the majority of mQTLs are shared between ethnic groups, our study also provides a comprehensive list of EA-specific mQTLs that may aid the understanding of the genetic architecture of human traits and diseases in this population. For example, in our revised manuscript, we demonstrate a trans-colocalization on chr21q22.2 that is significantly associated with basophil count in BBJ and also colocalizes with immune diseases such as urticaria, pericarditis, and asthma. This suggests that EA-specific mQTLs may play a crucial role in disease susceptibility in East Asian populations. Details on these findings can be found in our revised section "**East Asian-specific trans-colocalizations**".
3. We would like to emphasize the importance of functional annotation in comprehending the biological mechanisms that underlie genetic associations with human traits and diseases. Our East Asian mQTLs provide functional annotations for genetic susceptibility loci that are particularly significant in East Asian GWAS, but also for many loci found in other populations. For example, we provide evidence in our revised manuscript that a cis-colocalization at chr13q14.11 likely regulates its cis-associated mCpG and functionally influences the expression of *ELF1*, explaining the genetic association with adult height specific to East Asian populations. Further details on this can be found in our revised section "**East Asian mQTLs contribute to East Asian-specific genetic associations**". Additionally, we provide functional explanations for genetic associations between *FOSL2* and eosinophil count, and between *NFKB1* and ulcerative colitis (where BMI was used as an intermediate phenotype) from DNA methylation levels, and these genetic associations have been found in both Europeans and East Asians. For details, please refer to "**A FOSL2-mediated mQTL hotspot influences eosinophil counts**" and "**A NFKB1-mediated trans-mQTL hotspot may mediate the risk of obesity**".
4. From an evolutionary perspective, genetic and phenotypic differences between populations are the result of long-term genetic drift, mutation, and selection. Our study on East Asian populations and identification of EA-specific mQTLs might contribute to understanding the evolutionary history of human populations and the genetic basis of population-specific adaptations. In our revised manuscript, we found that East Asian-specific colocalizations were more likely to be trans than cis, and that these trans-colocalizations were more functional than cis-ones. Additionally, the associated trans-mQTLs show substantial allele frequency differences between East Asians and Europeans, indicating a potential relationship to evolutionary parameters such as drift, selection, admixture, or bottleneck. Our identification of EA-specific mQTLs and their functional effects provides an invaluable resource for future studies seeking insights into the evolutionary forces specific to this population. For details, please refer to our revised sections "**East Asian mQTLs contribute to East Asian-specific genetic associations**" and "**Cis- and trans-mQTLs impacting trait-associations via different patterns**".

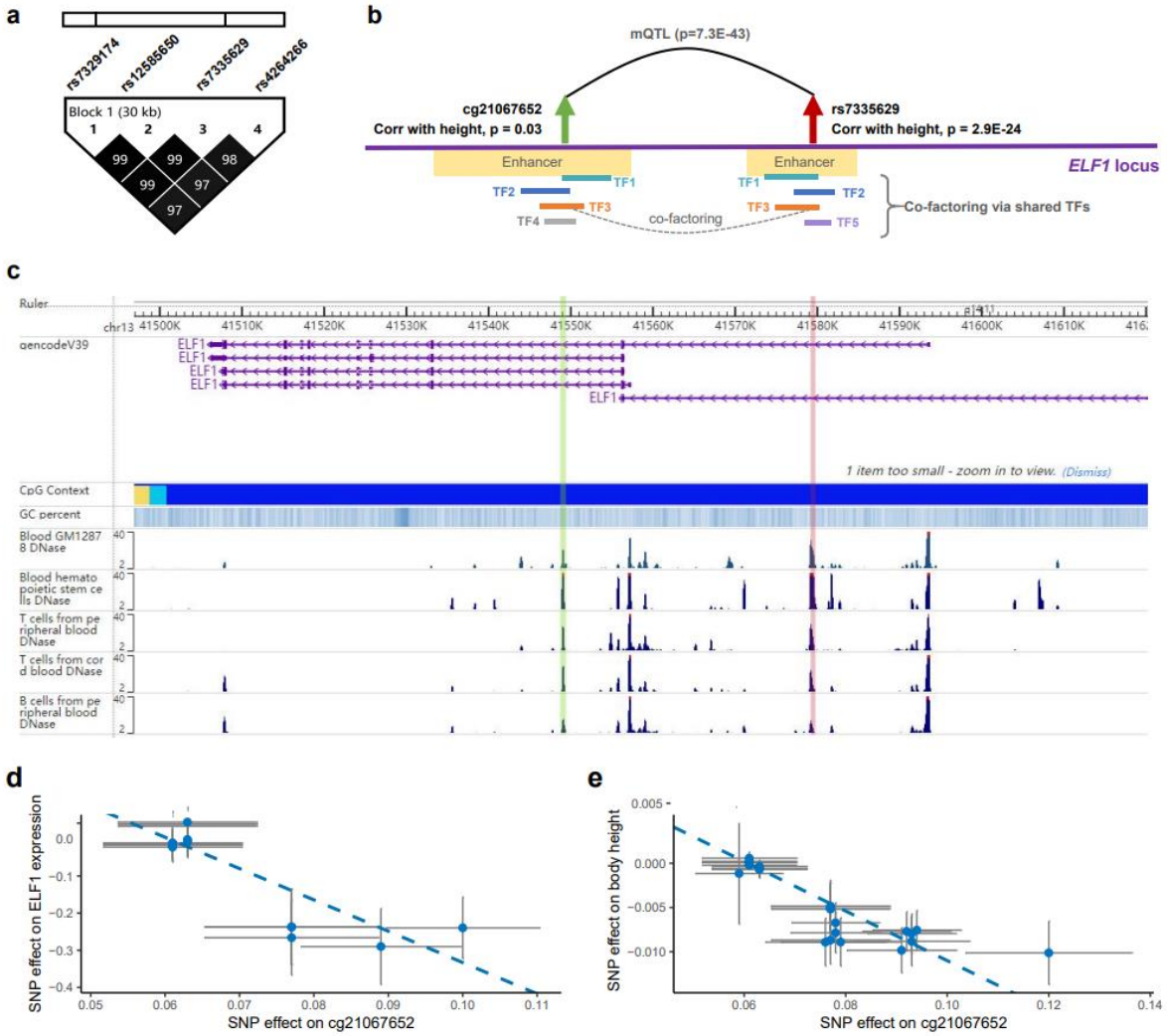
Comment-2. cis vs trans mQTLs. The bulk of the reviewer response talks about mQTL relationships without making the cis- trans- distinction. These are fundamentally different in terms of biological mechanism and interpretation. Trans-CpGs provide the opportunity to link remote loci into co-ordinate pathways. Many would argue that this opportunity to unravel trans-acting nuclear pathways has been the primary highlight from mQTL analyses. By comparison, the insights from a cis-mQTL are more modest (unclear even, see point 3). This limits

interpretation of the material; for example when discussing cell-specific mQTLs, is the evaluation of cis-mQTLs, trans-mQTLs or a mix (if so, what mix?). Similar for the relationships of the 3% of mQTLs that are EA 'specific' – cis, trans, and any inference from this?

Response: We appreciate the reviewer's suggestion and agree that distinguishing between cis- and trans-mQTLs is crucial for proper interpretation of our results. In response, we have extended our analysis by examining cis- and trans-mQTLs separately and have made several significant findings, as detailed in our revised sections "**East Asian mQTLs contribute to East Asian specific genetic associations**" and "**Cis- and trans-mQTLs impacting trait-associations via different patterns**". Thank you for pointing out this important issue.

More specifically, we revised our section "**East Asian mQTLs contribute to East Asian specific genetic associations**" on page 10 as "*Several variants of ELF1 have shown large effect on adult height in East Asian populations[2,3]. For instance, rs7335629 is reported to be an EA-specific signal in the latest human stature study[3]. This SNP and one of its associated CpG (cg21067652) are predicted to be in co-opening regions with binding sites from the same TFs (**Extended Fig.8b&c**). Furthermore, a two-sample MR analysis revealed a causal effect of cg21067652 on ELF1 expression and adult height (**Extended Fig.8d&e**). These results imply that colocalization of EA-specific mQTLs and trait GWASs could advance the understanding of biological mechanism of trait-associations from epigenetic level.*".

And a new **Extended Fig. 8** is added to demonstrate these results.



Extended Fig. 8 The cis-colocalization at chr13q14.11 provides epigenetic evidence for the East Asian-specific height-association (rs7335629-height).

a, The East Asian-specific height signal (rs7335629) is in high-linkage with three SNPs in the colocalization locus at chr13q14.11. **b,** rs7335629 has potential chromatin interaction with one of the CpGs (cg21067652) that colocalized at chr13q14.11. **c,** Both rs7335629 and cg21067652 are located in regions of high DNase in several blood cell lines. **d,** Two-sample MR result indicates that cg21067652 is a causal factor for ELF1 RNA expression. **e,** Two-sample MR result indicates that cg21067652 is a causal factor for height in East Asians.

We also added a new result section in our revised manuscript as bellow:

"***East Asian-specific trans-colocalizations.*** *Among all our trans-mQTLs (365K), mQTLs at 46 loci showed significant trans-colocalization signals ($P_{SMR} < 5.0\times10^{-9}$ & $P_{HEIDI} > 0.05$), i.e., they are associated with multiple independent CpGs (mean 16.28, SD 39.90) and are simultaneously associated with 23 distinct GWASs in BBJ. Of the 46 loci, 36 exhibited EA-specific trans-colocalizations with 486 independent mCpGs and 15 GWASs primarily associated with hematological traits (8 out of 15,* ***Table S8, Fig.2d****). The proportion of EA-specific*

trans-colocalizations was significantly higher than that of cis-colocalizations (odds ratio=4.48, 95%CI [2.05,10.65], ***Extended Fig.9a&9b****) compared to shared ones, and EA-specific trans-colocalization were significantly enriched in predicted transcriptional and enhancer regions (P-value < 5×10^{-8},* ***Extended Fig.9c&9d****), suggesting that trans-colocalization events are more population-specific and functionally significant than the cis-ones, as seen in our East Asian sample.*

The most significant EA-specific trans-colocalization was located in the intron of ERG on chr21q22.2 (rs80109907, in complete linkage with the index SNP rs77106233 in hotspot H16, ***Extended Fig.10a****,* ***Table 1****, $P_{SMR} = 7.9\times10^{-35}$,* ***Fig.2d&Fig.S6****), where the A allele was primarily positively (97%) trans-associated with 233 independent mCpGs ($8.6\times10^{-79} < P_{mQTL} < 4.8\times10^{-10}$) and positively associated with basophil count in BBJ ($P_{GWAS} = 1.1\times10^{-59}$). The A allele is common in East Asian populations (f = 0.11) but rare in European populations (f = 0.01,* ***Extended Fig.10b)****. A PheWAS analysis revealed significant albeit weaker associations of rs80109907 with several blood cell count in different populations (Japanese and Europeans), with the A allele being positively associated with eosinophils (P-value = 6.0×10^{-15}), red blood cells (P-value = 2.4×10^{-7}) and platelets (P-value = 2.2×10^{-4}), and negatively with neutrophils (P-value = 2.4×10^{-6}) and monocytes (P-value = 1.1×10^{-15}) (****Table S9, Extended Fig.10c****). A SMR analysis revealed weaker (compared to basophil count) but significant trans-colocalizations involving several blood cell count and immune-related diseases, e.g., monocytes ($P_{SMR} = 9.5\times10^{-9}$), eosinophils ($P_{SMR} = 6.7\times10^{-6}$), white blood cells ($P_{SMR} = 1.1\times10^{-4}$), urticaria ($P_{SMR} = 2.7\times10^{-3}$), pericarditis ($P_{SMR} = 3.2\times10^{-3}$), and asthma ($P_{SMR} = 3.6\times10^{-3}$) (****Table S10, Extended Fig.10d****). A two-sample MR analysis further identified 39 causal CpGs (FDR < 0.05) for these traits (reverse was not significant,* ***Extended Fig.10e****).*

Although these 233 CpGs were not significantly enriched in ERG binding sites, they were significantly enriched in motifs of 62 TFs (P-value < 5.3×10^{-5}, ***Table S11****), 13 of which (including RUNX1) were validated by blood cell ChIP-seq data[4,5] (****Table S12****). TAL1 and RUNX1, two of these TFs, interact directly with ERG and, together with ERG and four other proteins (LYL1, LMO2, GATA2, and FLI1), form a tightly combinatorial transcription factor heptad complex that plays an important role in the transcriptional regulation of hematopoietic stem cells[6,7], in which RUNX1 and LMO2 are specifically highly expressed in basophils[8,9] (****Fig.2e****). Protein interaction analysis revealed that these 62 TFs and ERG formed a large protein interaction network (****Fig.2f****). GO analysis showed that these TFs, together with the genes annotated at the mCpGs, were significantly enriched in biological processes related to hematopoiesis (P-value = 8.3×10^{-21}) and regulation of leukocyte differentiation (P-value = 1.5×10^{-16}) (****Fig.2g&Fig.S7****). These results support the view that trans-regulated DNA methylation changes affect the binding efficiency of multiple transcription factors in the ERG protein complex, further regulating the whole process of hematopoietic cell differentiation.*"

The Fig.2 is updated with the major findings of trans-colocalizations (**Fig.2d-2g**) and two extended figures (**Extended Fig.9 and Extended Fig.10**) is added to support the description of trans-colocalization findings.

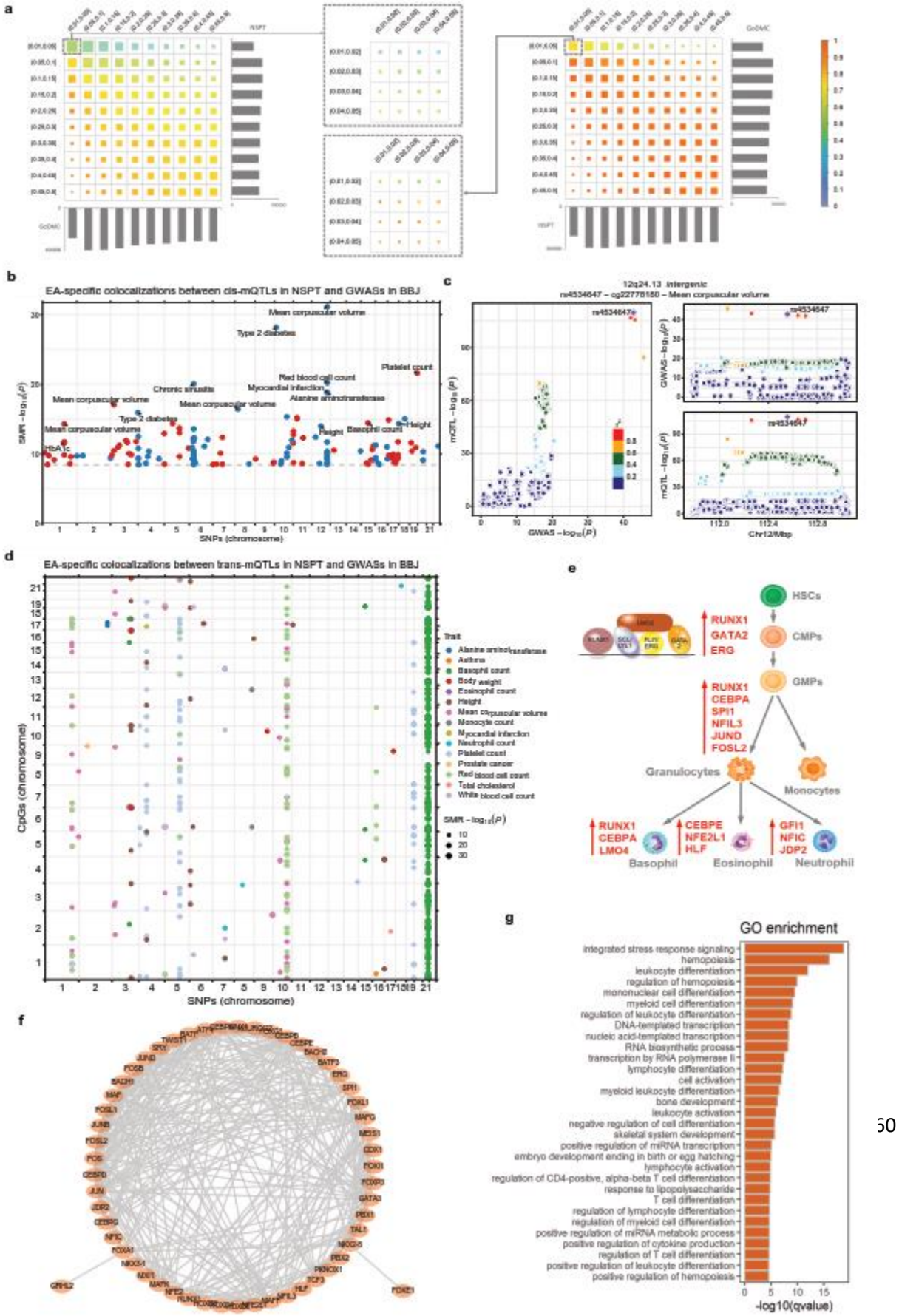
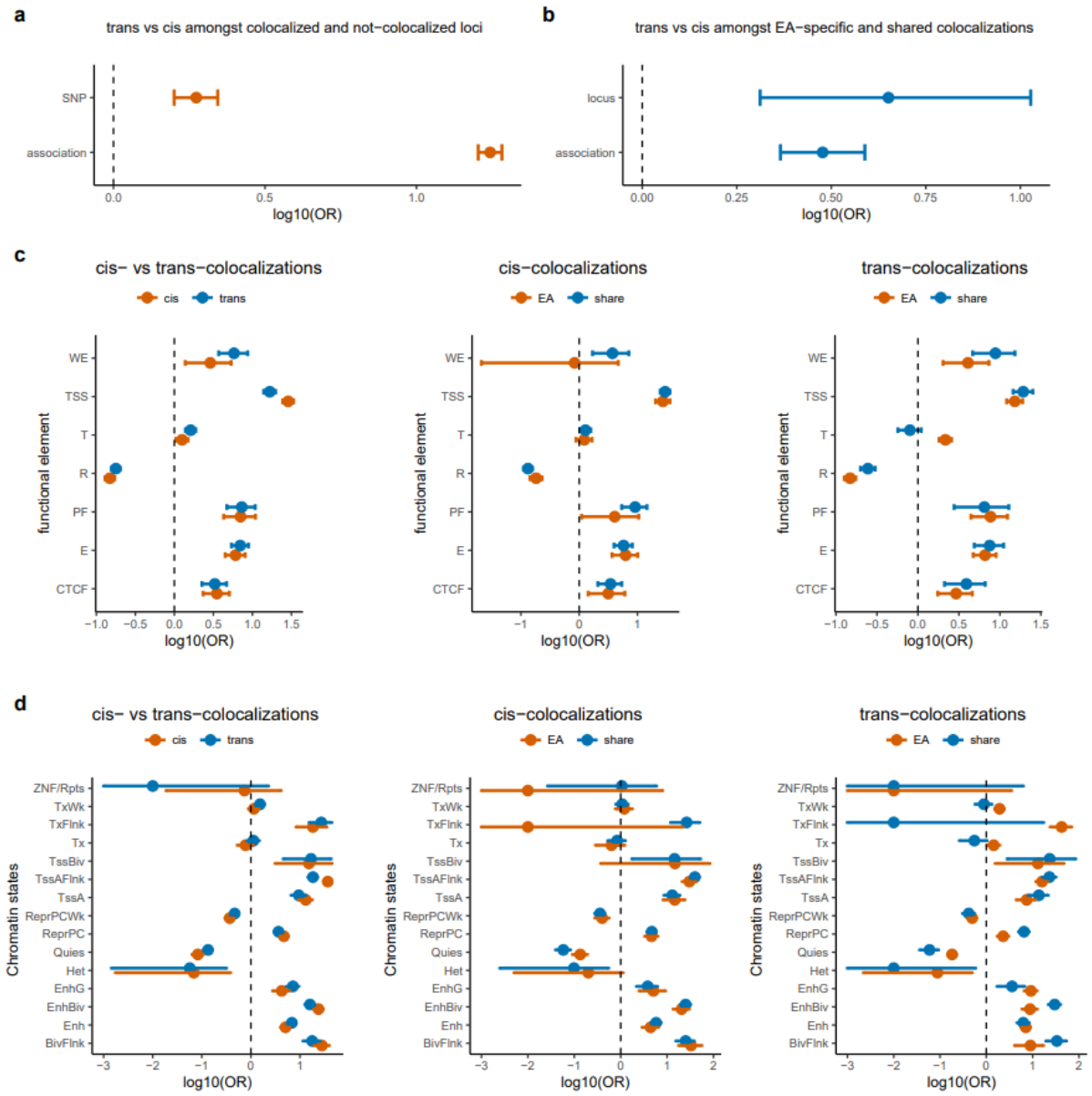

a
b
EA-specific colocalizations between cis-mQTLs in NSPT and GWASs in BBJ
SMR −log10(P)
SNPs (chromosome)
Mean corpuscular volume
Type 2 diabetes
Platelet count
Red blood cell count
Chronic sinusitis
Myocardial infarction
Alanine aminotransferase
Mean corpuscular volume
Mean corpuscular volume
Type 2 diabetes
Mean corpuscular volume
Height
Basophil count
Height
HbA1c
c
12q24.13 intergenic
rs4534647 – cg22778150 – Mean corpuscular volume
rs4534647
mQTL −log10(P)
GWAS −log10(P)
GWAS −log10(P)
mQTL −log10(P)
rs4534647
Chr12/Mbp
d
EA-specific colocalizations between trans-mQTLs in NSPT and GWASs in BBJ
CpGs (chromosome)
SNPs (chromosome)
Trait
Alanine aminotransferase
Asthma
Basophil count
Body weight
Eosinophil count
Height
Mean corpuscular volume
Monocyte count
Myocardial infarction
Neutrophil count
Platelet count
Prostate cancer
Red blood cell count
Total cholesterol
White blood cell count
SMR −log10(P)
10
20
30
e
HSCs
CMPs
GMPs
RUNX1
GATA2
ERG
RUNX1
CEBPA
SPI1
NFIL3
JUND
FOSL2
Granulocytes
Monocytes
RUNX1
CEBPA
LMO4
CEBPE
NFE2L1
HLF
GFI1
NFIC
JDP2
Basophil
Eosinophil
Neutrophil
f
g
GO enrichment
integrated stress response signaling
hemopoiesis
leukocyte differentiation
regulation of hemopoiesis
mononuclear cell differentiation
myeloid cell differentiation
regulation of leukocyte differentiation
DNA-templated transcription
nucleic acid-templated transcription
RNA biosynthetic process
transcription by RNA polymerase II
lymphocyte differentiation
cell activation
myeloid leukocyte differentiation
bone development
leukocyte activation
negative regulation of cell differentiation
skeletal system development
positive regulation of miRNA transcription
embryo development ending in birth or egg hatching
lymphocyte activation
regulation of CD4-positive, alpha-beta T cell differentiation
response to lipopolysaccharide
T cell differentiation
regulation of lymphocyte differentiation
regulation of myeloid cell differentiation
positive regulation of miRNA metabolic process
positive regulation of cytokine production
regulation of T cell differentiation
positive regulation of leukocyte differentiation
positive regulation of hemopoiesis
-log10(qvalue)

Fig. 2 the characteristics and potential applications of East Asian specific mQTLs

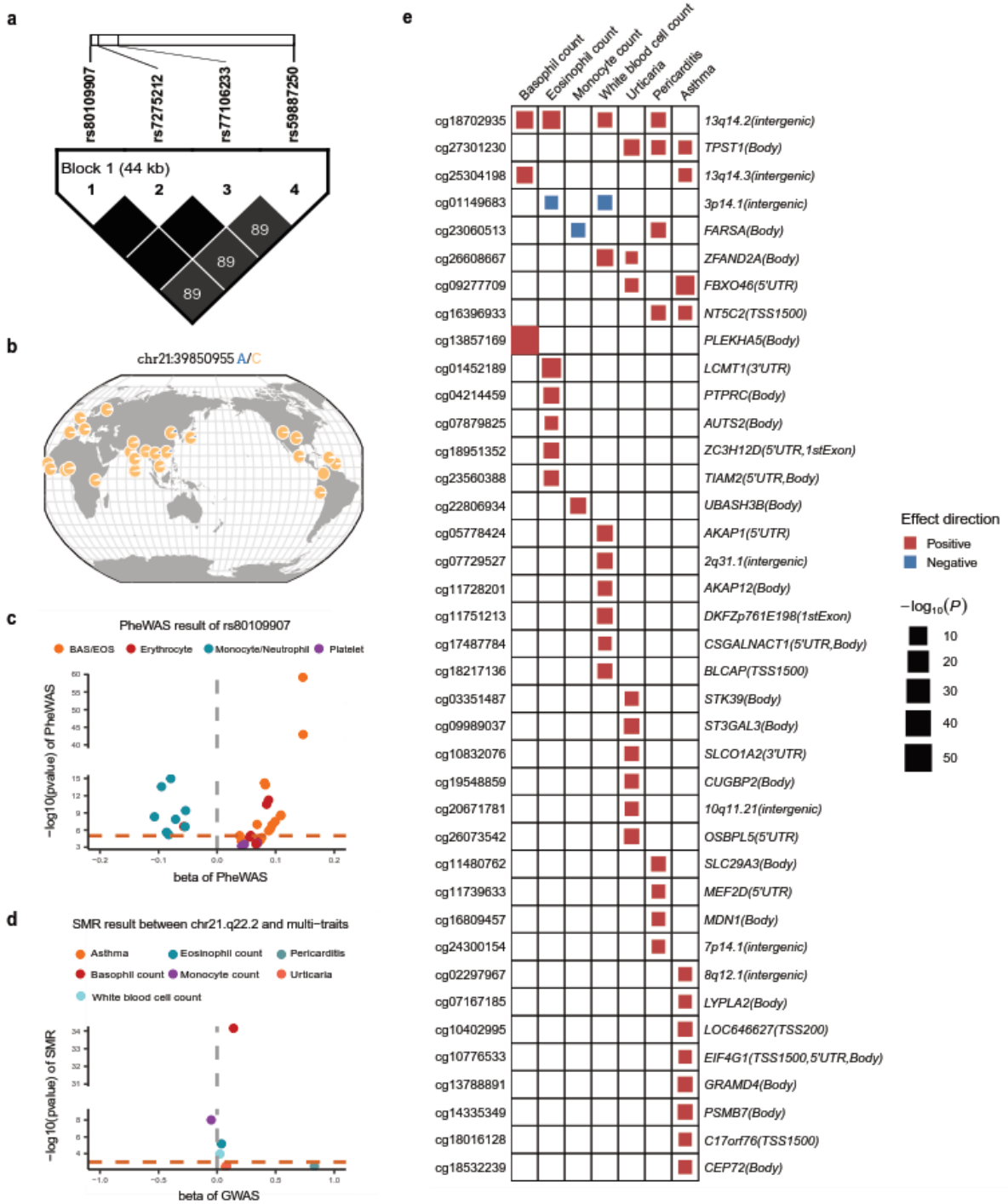
a, The left and right panels show the replication rates (redness) and numbers of mQTLs (square size) of study-wide significant mQTL in GoDMC and NSPT, respectively. The dependence on MAF in the later is particularly evident in the low MAF bins (middle panels). **b**, Manhattan plot showing 144 East Asian-specific cis-colocalizations (96 loci and 38 traits) with SMR significance on the y-axis and a Bonferroni-corrected threshold of $P < 3.5\times10\text{-}9$ indicated by the gray dashed line. **c**, The most significant East Asian-specific cis-colocalization on chr12q24.13, where an intergenic variant rs4534647 was cis-associated with cg22778180 in the first intron of MAPKAPK5 (lower right) and simultaneously associated with mean corpuscular volume (upper right). **d**, Scatter plot showing 541 East Asian-specific trans-colocalizations (36 loci and 15 traits), with point size representing SMR significance in East Asians. **e**, A schematic diagram of the hematopoietic tree showing the regulation of different genes during hematopoietic cell differentiation by the heptad complex comprising ERG, TAL1, RUNX1, LYL1, LMO2, GATA2 and FLI1. **f**, A protein-protein interaction network of ERG and 62 TFs enriched in the 233 CpGs trans-colocalized at chr21q22.2 loci (STRING). **g**, Enrichment of the chr21q22.2 trans-colocalization-related genes (ERG, 62 TFs and 195 annotated genes of trans-colocalized CpGs) in biological pathways and processes (Metascape).



Extended Fig. 9 The Enrichment of cis- and trans colocalizations in EA-specific colocalizations and functional elements.

a, Enrichment results of trans- vs cis-colocalizations amongst colocalized loci vs not-colocalized loci. **b**, Enrichment results of trans- vs cis-colocalizations amongst EA-specific vs EAS-EUR shared colocalizations. **c**, Enrichment results of cis- vs trans-colocalization loci in functional elements. Left: enrichment of cis- and trans-colocalization loci in functional elements; Middle, enrichment of East Asian-specific and EAS-EUR shared cis-colocalization loci in functional elements; Right, enrichment of East Asian-specific and EAS-EUR shared trans-colocalization loci in functional elements. **d**, Enrichment of cis- and trans-colocalization loci in chromatin states. Left, enrichment of cis- and trans-colocalization loci in chromatin states; Middle, enrichment of East Asian-specific and EAS-EUR shared cis-colocalization signals in chromatin states; Right, enrichment of East Asian-specific and EAS-

EUR shared trans-colocalization signals in chromatin states.



Extended Fig. 10 The relation between the trans-colocalization at chr21q22.2 and blood cell traits and

immune diseases.

a, The East Asian-specific colocalization signal (rs80109907) at chr21q22.2 is in complete-linkage with the index SNP (rs77106233) in trans-mQTL hotspot H16. **b**, The geographic distribution of rs80109907 allele frequencies in different populations (1000 Genomes Project) by the Geography of Genetic Variants (GGV) browser (https://popgen.uchicago.edu/ggv). **c**, The PheWAS result of rs80107709 (https://gwas.mrcieu.ac.uk/phewas). **d**, The colocalization result of chr21q22.2 with other blood cell count and immune-related diseases. **e**, Two-Sample MR results showing 39 CpGs are causal for 7 traits (several blood cell count and immune-related diseases) at FDR < 0.05.

Comment-3. Integrating mQTLs with GWAS. The authors report 98 loci at which a SNP associated with phenotype shows colocalization with mQTLs in EA but not EUR. On this occasion, the text does make clear the analysis is of cis-mQTLs. However, the results show that the 'ethnic specific effect' is simply determined by differences in allele frequency (the SNP is low frequency or absent in Europeans; Fig 2). How does the presence of a cis-mQTL finding advance understanding at these loci? Taking the ELF1 locus as an example, the association of the locus with height is demonstrated by the SNP (and supported by rare genetic variants) in EA, and the role of the gene is demonstrated by ELF as an eQTL. How does the presence of some co-localising cis-mQTLs enhance understanding of the ELF locus? What is the genomic mechanism or process revealed? Co-localisation may indicate a shared underlying genetic basis, but does not indicate whether methylation is cause, consequence or covarying with the phenotype. More importantly, given the low coverage of the EPIC array (~3% of CpGs), the analyses do not identify the specific 'functional' CpGs or the 'functional' SNP, or the potential genomic mechanism linking SNP to methylation. This extends the point that, while trans-mQTLs can reveal regulatory pathways, the presence of a cis-mQTL is currently less informative.

Response: We agree with the reviewer that the cis-colocalizations are more difficult to interpret due to the presence of LD. However, this is the same with eQTLs, which have shown tremendous utility in functional annotation of GWAS findings, i.e., the eQTLs are more likely to be functionally involved than the non-eQTLs, which is also true for our cis-mQTLs. We acknowledge the limitations of the EPIC array and the challenge of identifying the specific functional CpGs or SNPs in our analyses. However, we believe that identifying cis-mQTLs that colocalize with GWAS signals can still provide valuable information for functional annotation and understanding the underlying genetic basis of complex traits. We agree that co-localization does not indicate causality, but it can provide hypotheses for further investigation, such as identifying potential regulatory mechanisms that link SNP to methylation or exploring the role of methylation in the pathway leading to the phenotype. We have emphasized this point in **Discussion** in our revised manuscript and highlighted the need for further functional studies to validate our findings and uncover the underlying biological mechanisms, "*Colocalization of cis-mQTLs with GWASs can facilitate fine-mapping of trait-variants, as previously demonstrated*[10-18]*, whereas colocalization of trans-mQTLs with GWASs reveal biological pathways contributing to variant-trait association and the role of methylation in these pathways. We identified an EA-specific trans-colocalization that revealed 233 distant CpGs involved in basophil differentiation by affecting the binding efficiency of the ERG protein complex, providing an explanation for the difference between East Asian and European genetic associations at the DNA methylation level. However, future studies are needed to validate these findings.*".

As you have seen in our revised manuscript, we have made substantial additional efforts to find evidence that might reveal some biological mechanisms. We indeed have some very interesting findings, especially for the trans-mQTLs.

We have provided evidence that a cis-colocalization at chr13q14.11 likely regulates its-associated CpG and functionally influences the expression of *ELF1*, explaining the specific genetic association with adult height in East Asian populations. Details can be found in our revised section "**East Asian mQTLs contribute to East Asian specific genetic associations**".

We have shown that trans-colocalization events are more population-specific and functionally significant than the cis-ones, as seen in our East Asian sample. Details can be found in our revised section "**East Asian-specific trans-colocalizations**".

We have also provided an example of trans-colocalization at chr21q22.2 (intron of *ERG*) involving 233 trans-mCpGs and several hematological traits and immune-related diseases. It suggests that trans-regulated DNA methylation changes affect the binding efficiency of multiple transcription factors in the ERG protein complex, further regulating the whole process of hematopoietic cell differentiation. Details can be found in our revised section "**East Asian-specific trans-colocalizations**".

Comment-4. Cell specific effects (1). The issue about shared vs cell specific effects remains problematic. Hawe et al identify a set of mQTLs in whole blood, and show that many replicate in adipose tissue. It remains the case that the use of a mixed tissue such as whole blood for discovery will favour identification of mQTLs with low heterogeneity of effect between different cell types (ie that are shared). Respectfully, the reviewer was not confusing 'shared between ancestries' with 'shared between cell types'.
Response: The reviewer's original statement was that "*As one line of evidence of this they use the Hawe et al study which looks at cosmopolitan mQTLs from across populations and tests them in white cell subsets and isolated adipocytes. I would highlight that this test of generalizability in Hawe et al was limited to mQTLs selected to be shared. Without acknowledging this limitation, the authors are presenting what is a circular argument: things discovered as shared appear to be shared*". To us this statement does seem to imply confusion between 'shared between ancestries' with 'shared between cell types', because Hawe et al did not limit themselves to mQTLs selected to be shared. Given that myeloid cells make up 70-80% of immune-cells in blood, many of the mQTLs in Hawe et al could well be myeloid specific. So, we are glad that the reviewer has now clarified that there was no confusion.

In fact, let us now make it crystal clear that we agree with the reviewer "*that using CellDMC to infer mQTLs from a mixed tissue like blood favors the identification of cell-type independent mQTLs*". However, in contrast to the reviewer's statement, we opine that such a statement needs to be quantified. In other words, does our strategy of inferring cell-type specific mQTLs from blood hugely favor identification of shared mQTLs or does it only very mildly favor identification of shared mQTLs? The answer to this question depends on a multitude of factors, one being the threshold for calling original mQTLs in blood. As we have been arguing all along, and as demonstrated by the Monte-Carlo analysis in our last revision (following the reviewer's suggestion), by using a very relaxed P<1e-8 threshold when calling mQTLs, we are not missing a substantial number of cell-type specific mQTLs when performing the subsequent CellDMC analysis. Indeed, in the revised version of the manuscript, we performed all the analyses requested by the reviewer, whilst also improving our own previous analysis to provide a more rigorous derivation of the number of cell-type independent mQTLs in blood. These analyses clearly demonstrate that by using a very relaxed P<1e-8 threshold when calling mQTLs, that the subsequent cell-type specific mQTL analysis is not limited to shared mQTLs.

To re-emphasize these points: we have repeatedly acknowledged (and we have now done so again at the start of the Results section), that calling cell-type specific mQTLs from a mixed tissue like blood is a limitation, yet the reviewer offers no convincing or reasonable alternative to the in-silico computational approach adopted here, which we note is state-of-the-art [Zheng et al (2018)[19]] and which has led to the identification and successful validation of cell-type specific mQTLs, as clearly demonstrated in Fig.3a-e of our paper.

An important new addition to Fig.3, shown in the newly updated Fig.3g (displayed further below in response to another point), and which should help further alleviate the reviewer's concern, is the demonstration that myeloid and lymphoid specific mQTLs (mCpGs) are specifically enriched for CpGs that are significantly hypomethylated

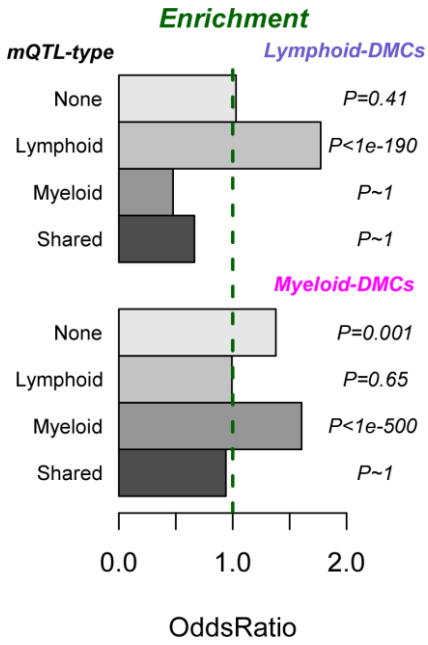
in the myeloid and lymphoid lineages, respectively. This provides substantial novel insight into the nature of lineage-specific mQTLs, namely that cell-lineage specific hypomethylated CpGs are more likely to represent mQTLs in the same lineage compared to other lineages. This now anchors our findings of myeloid and lymphoid specific mQTLs on concrete biology and serves as a form of indirect validation, which together with the direct validations in the external Han Chinese cohort and Blueprint's sorted data, makes a very strong case that our algorithm is successfully identifying cell-lineage specific mQTLs in blood.

Comment- 5a. Cell specific effects (2). I was initially unclear how the authors were defining a cell specific effect. The text states that 'many mQTLs significant in one-lineage also displayed associations in the other, albeit marginally so'. Their approach seems to be to define a cell-specifc mQTL as i. being present in one cell type at P<10-8, but ii. no evidence for association (failing to reach P<0.05) in the alternate cell subset(s) [is this any or all subsets?]. Based on these criteria, they conclude that mQTLs are shared across tissues. It is helpful that the approach has been clarified. The result aligns with population genetic studies (for example) which show that the great majority of genetic associations in one ancestral group can also be demonstrated in other ancestry at a permissive threshold of P<0.05.

Response: We are happy that the reviewer now understands our definition of cell-type specific mQTLs, and indeed most of our follow-up analyses were done at the resolution of two lineages (myeloid vs lymphoid). For the case of more than two blood cell-types, the definition of "specificity" becomes much more ambiguous because in theory an mQTL could be shared by only 2 of the 7 main blood cell-types, which some might consider "specific" whilst others would not. For this reason, and also because at the higher resolution of 7 blood cell subtypes we would require an even larger dataset to circumvent power-issues, we restricted the analyses at the resolution of two lineages (myeloid vs lymphoid).

Comment-5b. However, this approach fails to address the critical biological question – which of the mQTLs play a more important biological role in one cell subset that another? For example, is the functional consequence of genetic variant in NFKB1 the same in all cell subsets or is it greater in immune cell subsets, and if so which one? Such an analysis would provide real new insight into cell specific biology, analogous to the insights generated by cell specific studies of gene expression (eg GTEX). Instead, the present analysis adopts an approach that blurs the distinction between cell subsets, and creates the impression that mQTLs are the same across cell-subsets.

Response: We appreciate the reviewer's question and we have addressed this in a number of different ways. First, we have now extended the enrichment analysis shown in our original Fig.3g, to test enrichment of lineage-specific mQTLs among lineage-specific DMCs (see figure below). This demonstrates that myeloid-specific mQTLs are more likely to be DMCs that are hypomethylated in myeloid cells, and similarly that lymphoid specific mQTLs are more likely to be DMCs that are hypomethylated in lymphoid cells. This contrasts with the eFORGE2 analysis in old Fig.3g (now new Fig.3h), which had tested enrichment against immune-cell specific DHSs (as opposed to DMCs), and which had only demonstrated specificity for the myeloid-mQTLs. This is in line with the notion that cell-type specificity of mQTLs would correlate better with differential DNAm between lineages, compared to DHS. Thus, our new analysis nicely demonstrates potential functional significance of the myeloid and lymphoid specific mQTLs, and has been incorporated into panel Fig.3g:



Legend for Addendum to Fig.3g: *Enrichment analysis of four types of mQTLs (shared, myeloid-specific, lymphoid-specific, none) among lymphoid and myeloid DMCs. By "shared" we mean mQTLs that were called in both myeloid and lymphoid lineages. By "none" we mean mQTLs that were not significant in both myeloid and lymphoid compartments. Myeloid-DMCs are significantly hypomethylated (FDR<0.05) in myeloid cells compared to lymphoid cells, and conversely lymphoid-DMCs are significantly hypomethylated in lymphoid cells compared to myeloid cells. These DMCs were identified using EPIC DNAm data from Salas et al (2022)[20]. Odds Ratios and P-values derive from a one-tailed Fisher's exact test.*

Second, we should clarify that we had indeed explored whether the *NFKB1* trans-mQTL network was specific to an immune-cell subset. Because of the lengthy nature of this MS, we had not expanded on this point due to space restrictions, but our finding was that the trans-mQTLs associated with *NFKB1* were equally present in both myeloid and lymphoid compartments with no statistical evidence for a skew towards one particular lineage. This probably makes sense because we did not find *NFKB1* to display strong blood cell-subtype specific expression. We have now also applied mashr[21] to estimate the sharing of effect in lymphoid and myeloid in the two hotspots (*FOSL2* and *NFKB1*), which was 0.70 for *NFKB1* hotspot and 0.41 for *FOSL2* (**Fig. S21a&b**), suggesting that the *FOSL2* hotspot tend to have a cell-specific regulatory pattern. To gain insight into the biological significance of the hotspots, we applied functional enrichment analysis to the annotated genes of the hotspots. The results indicate that *FOSL2* and the annotated genes of *FOSL2*-mediated CpGs are in a highly cooperative protein-protein network (**Fig. S21c**), where the AP-1 protein family (FOSL2, JUND, etc.) serves as a major transcriptional regulator of MAPK1 and the latter is involved in a wide variety of cellular processes such as proliferation, differentiation, transcriptional regulation and development[22]. Remarkably, as shown in the main text, it also indicates that *FOSL2* and the annotated genes of *FOSL2*-mediated CpGs are enriched with traits of granulocytes (myeloid cells), especially eosinophil counts (**Fig. 6b**). These lines of evidence suggest that the SNP affects a phenotype involving complex network regulation, which may be fine-tuned by DNA methylation levels, and partly in a cell-specific manner, e.g. *FOSL2* hotspot and eosinophil count. We have added this results

in **Supplement** in the section "**15. Sensitivity analysis iv) validation of cell-lineage mQTLs"**.

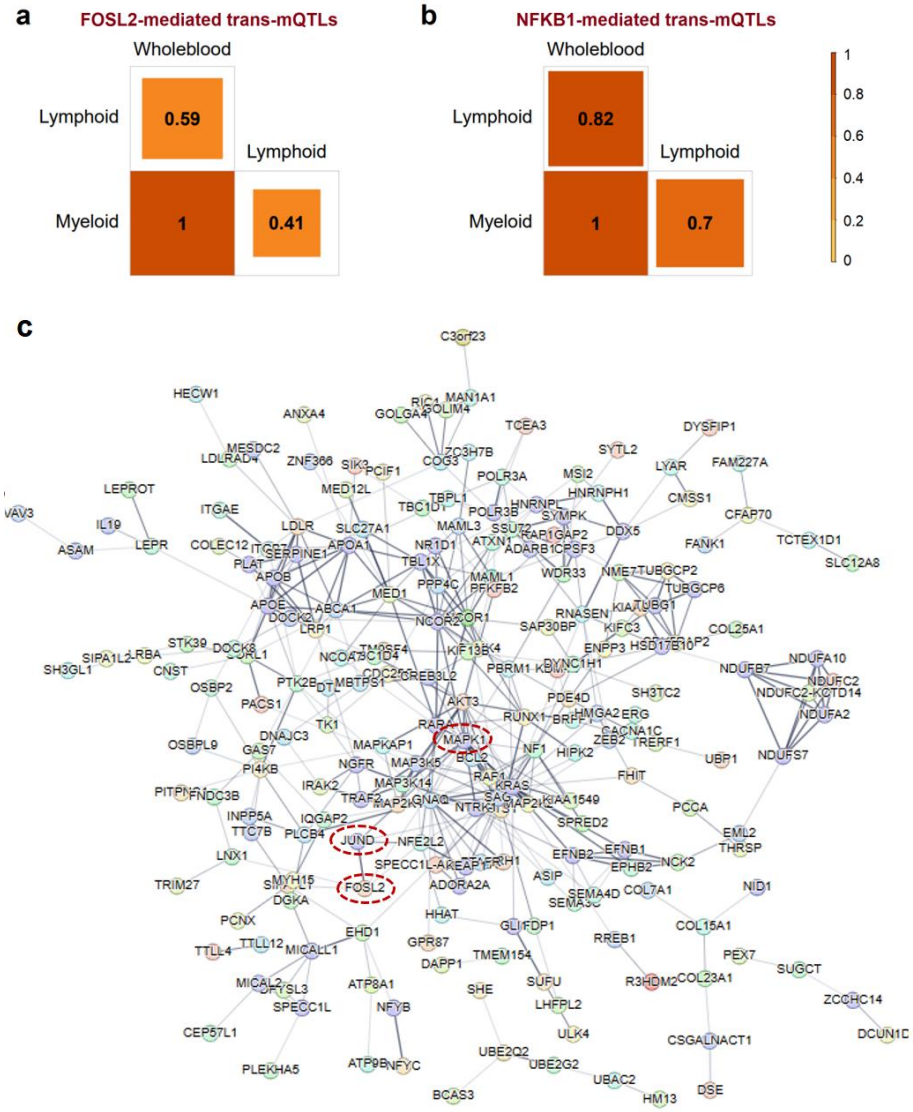


Fig. S21 FOSL2 and NFKB1-mediated trans-mQTLs in myeloid and lymphoid lineages.
a, Sharing effect of FOSL2-mediated mQTLs in whole blood, myeloid and lymphoid lineages; **b,** Sharing effect of NFKB1-mediated mQTLs in whole blood, myeloid and lymphoid lineages; **c,** The PPI network of FOSL2 and FOSL2-mediated genes (from STRING).

Third, the comparison with GTEX is in our opinion inappropriate. GTEX analyzed bulk tissue from different tissue-types and thus any insights gained should be termed "tissue-specific", not "cell-type specific", a critical distinction. Indeed, any derivation of "tissue specific" eQTLs from GTEX is subject to the huge caveat that there are differences in power between tissues (some tissues have many more samples than others) and that some tissues display much higher and variable levels of say immune or endothelial cell infiltration than others (e.g.

lung tissue may have over 40% immune-cells, whilst skin tissue would have far fewer). Moreover, the very recent mQTL study of eGTEX [Oliva et al (2022)[16]] concluded in the abstract that approximately only at least 5% of mQTLs display tissue-specificity. A glance at ExtendedDataFig.4b of that paper also reveals that the number of truly tissue-specific mQTLs is actually quite low for many of the tissues considered. Overall, that study suggests that the fraction of tissue-specific mQTLs could be in the range 5% to 30%. Given that the Oliva et al eGTEX study was very underpowered to detect shared mQTLs, the proportion of true tissue-specific mQTLs is more likely to be in the lower range of 5 to 15%, i.e. the fraction of shared mQTLs is likely to be quite high (85%-95%). Given that our study deals exclusively with one tissue (blood) and that it explores whether mQTLs are shared between major blood cell-lineages, it is therefore unsurprising to us that for blood tissue, the proportion of shared myeloid-lymphoid mQTLs is ~93%.

Finally, we would like to clarify to the reviewer that we have approached this question of cell-type specific mQTLs in blood with a very open mind, without falling into the trap of false preconceptions or prejudices. We think that the reviewer's opinion that "most mQTLs need to be cell-type specific" is not well founded. Our quantitative and rigorous analysis yields results that are entirely consistent with the estimates of cell-type independent mQTLs from 3 different studies: (i) Blueprint Chen L et al (2016) (study was underpowered but already concluded that at least 70-80% of mQTLs are independent of immune cell subset)[23]; (ii) Hawe et al (2022) (72-86% of their EU-SA mQTLs in blood validate in adipose cells, so at least this percentage is likely to be cell-type independent between more similar cell-types such as immune-cells in blood)[24] and (iii) Oliva et al (2022) (see above)[16]. As with any quantitative statements, we profoundly disagree with the reviewer's suggestion that the ~5% of immune cell type specific mQTLs are uninteresting. Indeed, as shown in our updated Fig.3g panel, myeloid and lymphoid specific mQTLs are more likely to be hypomethylated myeloid and lymphoid DMCs, respectively. The functional and biological significance of these will need to be explored in future studies.

Comment-6. New biology (1). Both shared CpG states at TADs and TF hotspots are recognized. In particular the 'hotspots' (eg compare Fig 5a with results in Bonder et al, Hawe et al). This is being oversold.
Response: We appreciate the reviewer's feedback and acknowledge that shared CpG states at TADs and TF hotspots have been previously described. However, we believe that our study adds novel insights to the field of cis-/trans-mQTL formation. For instance, we found that chromatin accessibility and transcription factor network could explain over 40% of cis-mQTLs. We also proposed that super enhancers may be responsible for the formation of trans-mQTL hotspots. In line with the functional role of super enhancers in cell identity, we detected several hotspots involved in hematological traits. One of them, located at chr21q22.2, contained a super enhancer and the *ERG* gene, where trans-regulated DNA methylation changes at 233 CpGs affect the binding efficiency of transcription factors in the ERG protein complex, further regulating the whole process of hematopoietic cell differentiation.

Regarding the *NFKB1*-trans-mQTL network, we believe it is important to demonstrate that this network is also present in East Asians. Assuming that this *NFKB1*-hotspot had not been present in EAs, this would then indicate that some of the differential genetic predisposition to ulcerative colitis between Caucasians and East Asians could be attributed to differential activation of the NFKB-pathway. That the *NFKB1*-hotspot is present in EAs, therefore, indicates that any differential genetic predisposition to the disease may be related to other pathways or factors. This is certainly important to help future studies direct their efforts and resources to evaluate what these other factors may be.

Comment-7. New biology (2). The trans-pathway linked to FOSL2 and blood counts is interesting, and would be compelling if supported by experimental evaluation of the statistical inferences. NFKB1 and BMI is less clear. If the NKFB1 SNP that drives the methylation in trans is not associated with BMI (line 1-2, page 28), then it is

difficult to understand how the data suggest that NFKB1 trans-CpGs are causal in obesity.

Response: With regards to *FOSL2*, we agree that experimental validation would provide additional support for our findings, but due to the tight time frame and limited budget we are unable to do this. We would like to note that our revised manuscript has become lengthy due to the additional analyses made in several rounds of extensive revisions, which further supports our findings. Nonetheless, we acknowledge that future studies should consider experimental validation to confirm the functional role of FOSL2 in eosinophil count regulation.

As far as *NFKB1* is concerned, we agree that the lack of association between the *NFKB1* SNP and BMI is not as anticipated, yet we also do not believe that this is contradictory. First of all, several EWAS have implicated the *NFKB1* pathway in BMI (Wahl et al (2017)- Chambers group)[25]. In line with this, we have found that many CpGs mapping to *NFKB1* binding sites are associated with BMI and are trans-mQTLs with the *NFKB1* index SNP. While the original Wahl et al study suggested that most BMI-CpGs are not causally implicated, a more recent study [Hawe et al (2022)[24]] by the same group concluded that a substantial fraction of BMI-CpGs could be causally implicated, highlighting the *UBASH3B* locus as one example. Our MR-analysis confirmed that most BMI-CpGs are a consequence of BMI but also suggested that the *NFKB1*-binding site trans-mQTLs associated with BMI could be causally implicated. It is therefore not contradictory that DNAm at these specific sites, which are only mildly influenced by the SNP acting in-cis on *NFKB1*, are also influenced by endogenous (other in-cis SNPs) and exogenous factors that may be causally linked to obesity. Although the subsequent interaction analysis we performed does not address causality it does nevertheless help identify further subsets where DNAm variation at the locus displays synergy between genotype and BMI (as a surrogate of exogenous factors such as diet). Indeed, it is worth highlighting again the fact that among the 3 CpGs exhibiting both causal and interaction effects, one mapped to *PTPN3*, a protein-tyrosine-phosphatase gene that has been linked causally to obesity [see Gurzov et al (2015)[26]], and another to *NOD2*, an intracellular innate immunity protein gene that has been shown to be protective of diet-induced obesity and colitis [Rodriguez-Nunez et al (2017), Gurses et al (2020), Kreuteret al (2019)][27-29]. Furthermore, we found that the genes annotated to *NFKB1*-mediated mCpGs display high cooperativity as a network centered on NFKB1 (**Fig. S22a**), and that these genes are enriched for metabolic and immune processes relevant for the regulation of BMI (**Fig. S22b**). Thus, overall, we think that our analysis focused on *NFKB1* binding site trans-mQTLs adds important insights to the current literature on this topic.

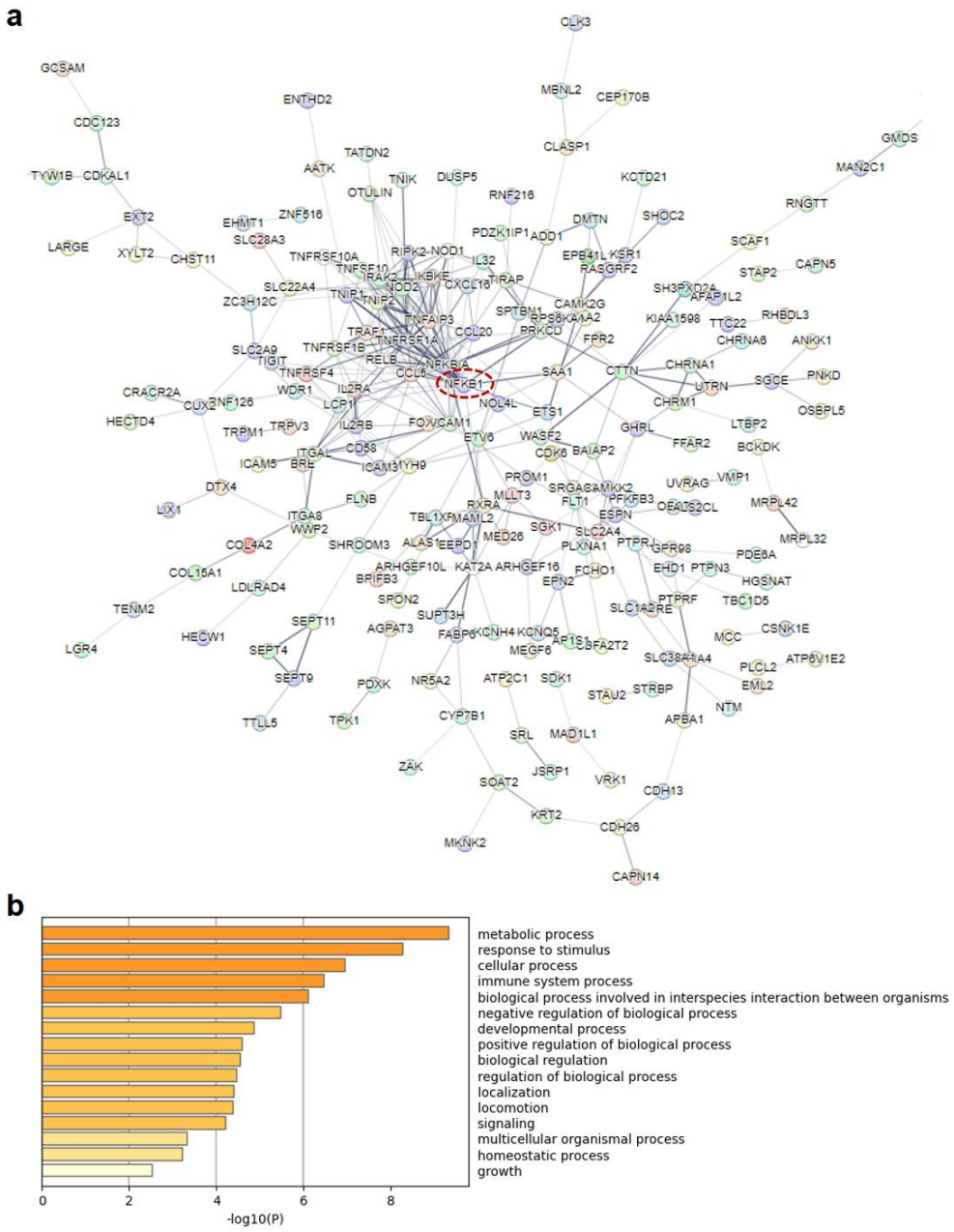


Fig. S22 The biological characteristics of NFKB1-mediated hotspots.

a, PPI network between NFKB1 and the annotated genes of NFKB1-mediated-mCpGs (STRING); **b,** The enrichment of NFKB1 and the annotated genes of NFKB1-mediated-mCpGs in GO (from metascape)

Reference:

1. Breeze, C.E., Beck, S., Berndt, S.I. & Franceschini, N. The missing diversity in human epigenomic studies. *Nat Genet*

54, 737-739 (2022).
2. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nat Commun* **10**, 4393 (2019).
3. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704-712 (2022).
4. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res* **50**, W175-82 (2022).
5. Oki, S. *et al.* ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO reports* **19**(2018).
6. Wilson, N.K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-44 (2010).
7. Hoang, T., Lambert, J.A. & Martin, R. SCL/TAL1 in Hematopoiesis and Cellular Reprogramming. *Curr Top Dev Biol* **118**, 163-204 (2016).
8. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663-77 (2015).
9. Hamey, F.K. *et al.* Single-cell molecular profiling provides a high-resolution map of basophil and mast cell development. *Allergy* **76**, 1731-1742 (2021).
10. Morrow, J.D. *et al.* Human Lung DNA Methylation Quantitative Trait Loci Colocalize with Chronic Obstructive Pulmonary Disease Genome-Wide Association Loci. *Am J Respir Crit Care Med* **197**, 1275-1284 (2018).
11. Taylor, D.L. *et al.* Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc Natl Acad Sci U S A* **116**, 10883-10888 (2019).
12. Huan, T. *et al.* Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. *Nat Commun* **10**, 4267 (2019).
13. Zhao, T., Hu, Y., Zang, T. & Wang, Y. Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. *Front Genet* **10**, 1021 (2019).
14. Min, J.L. *et al.* Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet* **53**, 1311-1321 (2021).
15. Soliai, M.M. *et al.* Multi-omics colocalization with genome-wide association studies reveals a context-specific genetic mechanism at a childhood onset asthma risk locus. *Genome Med* **13**, 157 (2021).
16. Oliva, M. *et al.* DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nat Genet* **55**, 112-122 (2023).
17. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat Neurosci* **19**, 48-54 (2016).
18. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**, 918 (2018).
19. Zheng, S.C., Breeze, C.E., Beck, S. & Teschendorff, A.E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods* **15**, 1059-1066 (2018).
20. Salas, L.A. *et al.* Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun* **13**, 761 (2022).
21. Urbut, S.M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187-195 (2019).
22. Karin, M. The regulation of AP-1 activity by mitogen-activated protein kinases. *J Biol Chem* **270**, 16483-6 (1995).
23. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
24. Hawe, J.S. *et al.* Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. *Nat Genet* **54**, 18-29 (2022).
25. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81-86 (2017).
26. Gurzov, E.N., Stanley, W.J., Brodnicki, T.C. & Thomas, H.E. Protein tyrosine phosphatases: molecular switches in metabolism and diabetes. *Trends Endocrinol Metab* **26**, 30-9 (2015).
27. Rodriguez-Nunez, I. *et al.* Nod2 and Nod2-regulated microbiota protect BALB/c mice from diet-induced obesity and metabolic dysfunction. *Sci Rep* **7**, 548 (2017).
28. Gurses, S.A. *et al.* Nod2 protects mice from inflammation and obesity-dependent liver cancer. *Sci Rep* **10**, 20519 (2020).
29. Kreuter, R., Wankell, M., Ahlenstiel, G. & Hebbard, L. The role of obesity in inflammatory bowel disease. *Biochim*

Biophys Acta Mol Basis Dis **1865**, 63-72 (2019).

Decision Letter, second revision:

7th Apr 2023

Dear Sijia,

Thank you for submitting your revised manuscript "Comprehensive mechanistic characterization of mQTLs in an East Asian population" (NG-A58885R3). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

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

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

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

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087

Reviewer #1 (Remarks to the Author):

Manuscript is further improved. No additional comments.

Final Decision Letter:

2nd Aug 2023

Dear Sijia,

I am delighted to say that your manuscript "Analysis of blood methylation quantitative trait loci in East Asians identifies ancestry-specific effects associated with complex trait variation" has been accepted for publication in an upcoming issue of Nature Genetics.

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

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

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

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

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office (press@nature.com) after sending your e-proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number (NG-A58885R4) and the name of the journal, which they will need when they contact our Press Office.

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

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

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

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where

possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including <a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Please note that Nature Portfolio offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

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

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

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

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

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

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

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, natureprotocols.com. If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in natureprotocols.com, you are enabling researchers to more readily reproduce or adapt the methodology you use. Natureprotocols.com is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to https://protocolexchange.researchsquare.com/. After entering your nature.com username and password you will need to enter your manuscript number (NG-A58885R4). Further information can be found at https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#protocols

Sincerely,

Michael Fletcher, PhD
Senior Editor, Nature Genetics

ORCiD: 0000-0003-1589-7087