nature research

Peer Review Information

Journal: Nature Genetics
Manuscript Title: Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation
Corresponding author name(s): Dr Josine Min

Editorial Notes:

Transferred manuscripts	This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Genetics.

Reviewer Comments & Decisions:

Decision Letter, initial version:

17th Nov 2020

Dear Dr Min,

Your Article, "Genomic and phenomic insights from an atlas of genetic effects on DNA methylation" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

As you will see, Reviewer #1 thinks that the manuscript has improved and is mainly happy with the revised version. There are a few minor comments that should be addressed, but these seem very doable.
Reviewer #2 also is content with the majority of the responses. There are some additional comments and requests for inclusion of more discussion about the limitations and caveats. While we recognize that you have done a lot of this already, we think that these are fair points that should be taken on board.
Reviewer #3 is satisfied with the revision.

Altogether, this has been determined to be technically sound and a great resource, so we ask that you carefully address these final points before we proceed.

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.

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

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

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on

the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

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

Thank you very much.

All the best,

Catherine

--

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071

Referee expertise:

Referee #1: QTLs, computational biology

Referee #2: epigenomics

Referee #3: QTLs, omics data

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
I believe that the authors have addressed all of the major concerns from my initial review. I really like the simulation study that you have performed for the colocalisation analysis. It address the issues that I had about potential false positives in the analysis. Minor comment: On line 681-682 of the supplementary data you write that "… we use the sparse approach for colocalisation analysis when **incomplete** regional data around the mQTL is **unavailable**." This double negative is a bit hard to parse, did you mean complete instead of incomplete?

I also appreciate that you have added the caveat to the discussion that "Fourth, DNAm might be on the causal pathway in a disease-relevant cell type or context.". This addresses my concerns about

potential cell type specific effects. I agree that it is currently not feasible to systematically evaluate this possibility due to the lack of appropriate cell-type-specific data. I also agree that your result about meQTLs in whole blood not generally being on the causal path for complex traits is still a very important finding, especially in the context of EWAS studies that are often based on the same tissue and DNA methylation array. I would definitely recommend all future EWAS studies to very carefully consider your results. I also believe that many of the same issues might be encountered when trying to use bulk tissue trans-eQTLs in causal inference.

Below are two minor comments:

I have now tried to read the two-dimensional enrichment analysis multiple times, but I am still not sure I understand how to interpret the results. Let's look at the IRF1 example. As I understand it right now, you have found that DNA methylation sites near the binding sites of the IRF1 TF are enriched to have trans-meQTL SNPs near the binding sites of EZH2, SMC3, ATF3, BCL3, TR4 and MAX TFs. Since you also mention that these trans-meQTLs could be driven by TF expression (line 467), does this imply that the binding of EZH2, SMC3, ATF3, BCL3, TR4 and MAX factors at these distal sites somehow influences the expression (or activity) IRF1 via some uncharacterised trans-acting factors? And then the change in IRF1 expression is what ultimately influences the DNA methylation of the target sites? I believe the analysis is sound so I don't think this issue should hold up the publication of this study, I was just struggling to understand what the interpretation of the results would be.

I am still not convinced that it is meaningful to split trans-meQTLs into cis+trans (covering also cis-close-to-trans) and trans-only categories, but since this distinction does not significantly change the major conclusions of the manuscript, I accept the authors decision to keep it as it is. However, there's one more thing that I would like the authors to check: are the trans-only meQTLs significantly further away from 450k probes than cis+trans and cis variants? My reasoning is that maybe trans-only meQTLs do not have any cis associations just because they happen to be far away from the sparse set of DNA methylation sites profiled by 450k arrays (i.e. they are near distal regulatory elements not captured by 450k arrays). Thus, perhaps you would also find cis associations for these if your were to profile the DNA methylation using a technology that has higher coverage of the genome. This would be easy to check by comparing the distribution of distances from the lead meQTL variant to the nearest 450k probe included in your analysis for the three meQTL categories.

Reviewer #2:
Remarks to the Author:
Responses to Min et al. replies to
Referee #5 (Remarks to the Author):
>>> Indicates Reviewer Reply

Min et al. have performed a multiple European ancestry cohort mQTL meta-analysis (27k with 5k replication) from 36 published cohorts in DNA derived from peripheral and cord blood. This analysis was performed with the previous generation Illumina DNA methylation 450k array. ~10M genotyped SNV were used (imputed 1000G panel). From 420k post-QC CpG probes they identify that ~45% (190k) of all the DNA methylation sites on the array are influenced genetically by identified mQTL – totalling >270k independent mQTL. Due to the power of this study, this was almost a 2x increase in

cis- and 10x in trans associations, respectively, compared to a similar study on one of these cohorts by Bonder et al. This later trans finding included 23k independent mQTLs (8.5% of total). This expansion in trans findings - with the high statistical threshold ($p<1e-14$) required for these due to the large testing space - is a highlighted strength of the meta-analysis. These mQTL in total are seen to explain 15-17% of the additive genetic variance of DNAm. Well-acknowledged statements are made regarding the genetic influence on DNAm levels being highly polygenic that clearly fit with previous experimental and genome-wide analyses.
Whilst shared genetic factors associated with both blood DNAm levels and complex diseases are identified, no strong casual relationships from DNAm to trait are supported – again in line with many current common DNAm EWAS analysis interpretations. The analysis of the expanded trans results are further extended into networks, although without any functional evaluation to further support these. Selective findings state that DNAm exhibits signatures of negative and positive natural selection, but do not provide evidence that the epigenetic change is itself causally driving this. The use of the now superseded 450k array clearly enables large numbers of samples from multiple cohorts to be included in this metanalysis. >90% of these probes are still included on the later EPIC array. Also, it is still useful for generalisable results as long as the bias of this array is acknowledged. Additionally, as the majority of EWAS to date have been performed with this 450k platform, enabling further reinterpretation of these previous studies' conclusions which could be very beneficial. Similar work has been performed in some of these cohorts individually before and this extends on papers as mentioned from Bonder et al. Nature Genetics 2017, as well as Blueprint consortium Chen et al Cell 2016, etc. Whilst, older papers can be used to justify points made, instead, it would be good to see that these data can convincingly lay down the state of current knowledge. Referring to DNAm as a complex trait does make for some chunky language at times and can lead the reader to inflate the potential functional significance of every detected mQTL.
• After careful checking we now changed two sentences that refer to DNAm as complex traits.
We changed
"As with classical complex traits31, much larger sample sizes are required to map associations involving small genetic effects in order to permit greater understanding of the genetic architecture and the biological processes underlying DNAm32."
to
"Much larger sample sizes are required to map associations involving small genetic effects in order to permit greater understanding of the genetic architecture and the biological processes underlying DNAm32."
and
"This revealed a genetic architecture of DNAm that is polygenic, as with many traits traditionally considered 'complex'."
to
"This revealed a genetic architecture of DNAm that is polygenic."
>>> Good to acknowledge and improve this

A recurring issue detailed in various points below is that the authors need to keep in the forefront of all of their conclusions that they are only testing 1.5% (420k of 28M CpGs) of the DNA methylome. Whilst statistically some short-range co-methylation can be detected, the limited array findings do not benefit anywhere near to the same level as genetic arrays do via LD in capturing common variation. The implications of this from the current findings – causality, variance, heritability, cis+trans identification, selection analyses etc need to be acknowledged.
• We apologise that the wording in the text may have strayed from emphasising that the 450k array only represents a small proportion of the methylome. This is something that we are keenly aware of

and highlight the point in the Implications section at the end of the paper. However, as the reviewer rightly points out, these are the sites that have been under analysis by the field, and our paper only seeks to evaluate the relevance of these sites. The extent to which this speaks to the total relevance of DNAm genome-wide remains to be seen, but it must be noted this paper identifies the largest number mQTL ever discovered on a much larger number of participants and studies than in previous efforts. The analysed sites represent 99% of RefSeq genes with multiple probes per gene, 96% of CpG islands from the UCSC database, CpG island shores and shelves, FANTOM4 promoters, the MHC region and some enhancer regions and additional content1 (so, to be fair they are more like extended exome arrays than genome-wide arrays if a comparison to SNP chips must be made). Although, underrepresented in distal regulatory elements (i.e. enhancers) and representing only 1.5% of all DNAm sites across the genome, these sites are actually quite representative of the known functional methylome. We agree that extrapolating our findings beyond the measured sites is not possible in this study. Rather, the purpose of this study is to evaluate the properties of the sites that have been routinely analysed in hundred thousands of datasets for the past decade (101237 datasets uploaded in Gene Expression Omnibus as "GPL13534"), in EWAS studies and for reference genomes including BLUEPRINT3, International Human Epigenome Consortium4 and The Cancer Genome Atlas consortium (https://tcga-data.nci.nih.gov/tcga/). We would like to acknowledge that profiling the entire methylome would also not be sufficient to explain the underlying complex genotype-methylome map as the CpGs will have variable genetic and environmental contributions. A whole methylome sequencing study of twins5 demonstrated that only a small proportion (~15–20% of 28M CpGs of the methylome) is differentially methylated in the population and that non-shared unique environmental factors are the main source of variation.

>>> As stated, the major and useful benefit of this current consortium analysis is to be able to re-examine the large number of previous studies that have been performed with this now superseded array platform. However, the additional statements by the authors that this 450k array captures an equivalent portion of epigenomic functional information as the almost complete coding genetic information gathered from an "extended exome array" conveys a poor understanding of the epigenome. This array DNA methylation array is not, as the authors state, "actually quite representative of the known functional methylome" – it is was designed a decade ago and was focused on promoters and genic regions. Beyond the fact that this array only covers only 1.5% of the total CpGs, all of our accumulated knowledge of the DNA methylome since this time (ENCODE, Roadmap etc) has identified that promoter regions, being CpG dense, are highly invariant. They remain hypomethylated across all tissue types, making many of these CpGs on the array redundant. Furthermore, genic regions are not found be to be enriched for other regulatory information, instead this predominately resides beyond in distal non-coding regions. The array does include coverage of a representation of CpGs that reside in more dynamic intermediate density CpG shore regions, but lacks any of the focused coverage of currently known enhancer regions included in its successor, the EPIC 850k array (EPIC itself still only covers 58 % of FANTOM5 enhancers, and 7 % distal and 27 % proximal ENCODE regulatory elements [Pidsley et al., Genome Biol. 2016]). These enhancers regions are considerably more dynamic and significant in term of driving functional lineage changes etc [Heintzman at al., Nature 2009, Ziller et al, Nature 2013, etc]. Additionally, all current methodologies poorly represent the ~50% of CpGs that reside within repetitive genome. Increasingly accumulating evidence indicate that these loci play a significant pathogenic role in disease-related regulation [Chuong et al., Nat Rev Genet. 2016] by perturbing transcriptional regulatory networks. The authors need to be more circumspect about what they are actually assaying and its limitations.

>>> It is unclear what point to help justify their meta-analysis array study that the authors are citing Busche et al. - a low coverage whole genome bisulphite methylome sequencing study of 30 twins - for? Clearly looking at a larger fraction of the DNA methylome would identify more genetic effects.

Firstly, even that study identified ~15–20% CpGs were differentially methylated, which is in fact still ~4-6 million CpGs, so not a "small" as the authors state. Secondly, that study's coverage was <10x, therefore, significantly less than the Roadmap standard of 30x. Its conclusions are limited to the resolution that that level of coverage can convey. Detection of small percentage DNA methylation changes, which robust array studies have shown us are the predominate observations in the DNA methylome, are required to interrogate fully common factors; such as genetic influence and cell-type heterogeneity. Furthermore, saturation analysis has shown that in fact ~100x is required for robustly differential methylation analysis at individual CpGs [Libertini et al., Nat Biotech. 2016]. Without robust high-resolution data, delineating the proportion of genetic to non-genetic effects cannot be concluded. In fact, more accurate array studies in monozygotic twins have found the reverse, by removing the large genetic component leading to almost a complete loss of identifiable disease-related DMPs compared to population studies [Paul et al., Nature Comms 2016]. It is the robustness of the CpGs probed, although limited in number, by these Illumina DNA methylation arrays that is the clear advantage of this methodology over low-coverage bisulphite sequencing. Subsequently, conclusions may be able to be drawn, albeit for these small number of CpGs, with potential broader implications for the complex genotype-methylome map.

• We have moved the analysis on the array and the mQTL coverage from the Implications to the Results to clarify this issue.
>>> Good to acknowledge and clarify this

Clearly the scale of this meta-analysis is a major undertaking, for which the authors should be commended in pulling together. It would, therefore, be good to further integrate the results with the latest functional understanding of the DNA methylome, where feasible, whilst clearly acknowledging any potential weaknesses due to lack of data from all of the DNA methylome. The authors state that they have identified a more complex genotype-phenotype map than has previously been hypothesised. These insights could be more strongly stated in the Abstract and Conclusion, if possible, particularly in regard to EWAS interpretation of previous studies. Further discussion of the responses to Reviewer 2 points as well as some additional comments are detailed below for the authors to consider.
• We thank the reviewer for this comment and we now changed the Discussion. We didn't change the abstract due to the word count.
"Fourth, DNAm might be on the causal pathway in a disease-relevant cell type or context.
…
Finally, EPIC arrays2 and promising low-cost sequencing technologies19 offer an opportunity to make detailed interrogations of enhancer and other regulatory regions, which may be more fruitful in finding causal relationships with complex disease."
…
"Overall our data and results have resulted in the most comprehensive atlas of genetic effects to date. We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."
>>> Good to acknowledge and improve this

Points from Responses to Reviewer #2
In regard to definitions used here for cis and trans – consistency with previous publications (GTex etc) is a fair explanation and a good approach. However, the definition of trans does lead to a less clear and the more overly descriptive nature of the results with the potential dilution of 'true' trans effects by cis. So, the biological interpretation of the 80% of trans that are inter-chromosomal is clearly more

straightforward. That "cis and trans mQTL operate through distinct mechanisms" (line 337) also fits with known experimental work from Stadler et al., Leinert et al., and others and this has been well conveyed by previous genome-wide publications, such as the Bonder et al. Paper.
• There is a distinction to be made between trans effects that may be biologically cis because they are on the same chromosome, versus trans variants that are proximal to other variants that influence DNAm in cis, or are themselves additionally influencing DNAm in cis. For the first case, we evaluate the extent to which omitting intra-chromosomal trans-effects are driving any enrichments and find that they are not. For the second case, we separate variants into categories that include 'trans-only' - in order to ensure that the trans-signals in enrichment analyses are not contaminated by potential cis-signals. This is something that previous studies have not done and has led them to make conclusions about trans operations that we can show are potentially unreliable.
>>> Fair enough

• We clarified this in the manuscript: "Our analysis shows that trans-only sites and SNPs have different properties as cis+trans SNPs and sites, indicating that enrichments of general trans categories are dominated by their cis functionality."
>>> Ok

It would be helpful if the authors could more clearly convey to the reader the mechanism of the cis effects observed and what this implies for their potential functionality. The DNA methylome is encoded via CpG Density and TFBS motifs – and cell-specific utilisation of this baseline state (as in the cited Schubeler review). Population allelic variation in both these genetic factors is a clear mechanism for the majority of cis effects described, as well as captured by variation in LD with these effects (Greally, Current Opinion Systems Biology 2017). This also encapsulates simple proximity CpG Density effects (Grandi et al. Genome Research 2015) and sequence-dependent Allele-Specific Methylation (Kerkel et al. Nature Genetics 2008). The genetic effects observed here were, as expected, identified to be highly stable across cohorts. As Figure S3a shows the mQTL and CpG couple, on the whole, reside within very close proximity of each other (This figure should potentially be included in the main paper to help convey this). The large impact of genetic effects on population EWAS has been clear from comparison with Monozygotic Twin EWAS that led to almost complete loss of phenotype-related DNAm findings (e.g. Paul et al. Nature Comms 2016). So, the high level of genetic effects identified here on DNAm in this meta-analysis is not surprising. Chen et al. (Cell 2016) identified considerable genetic influences even in isolated immune cell epigenomes and specifically urged caution in EWAS interpretation because of this – which the authors could further explore as detailed below.
• We agree with the reviewer that other possible cis mechanism for genetic variation exists such as sequence-dependent allele specific DNA methylation or structural variation. Here, we focused on population allelic variation captured by genotype and DNAm arrays which is of interest for EWAS and other population-based studies. Future studies using long-read sequencing where someone can determine both the DNAm level and the genotype from a single
molecule could reveal biological insights of these mechanisms (allele specific methylation and structural variation).
>>> Including discussion of this beyond just reporting associations would be informative for those readers from the non-epigenetic community.

• In our paper we explain this mechanism in the Introduction: "Because genetic influences on DNAm in blood have been shown to be widespread33,34, a powerful avenue into researching the functional consequences of changes in DNAm levels is to map genetic differences associated with population-level variation, identifying DNA methylation quantitative trait loci, (mQTL) that include both local (cis

mQTL) and distal (trans mQTL) effects."
>>> Ok

Causality and DNA methylation - DNAm sites have shared causal variants with complex traits via colocalization analysis - but the authors only have data on 1.5% of the DNA methylome. Therefore, the number of DNAm sites varying with casual variants is significantly larger. The authors state that "emerging ideas that DNAm variation is a marker of regulatory states and not a primary causal driver" – however, isn't this an acknowledged idea in the field?
• There is considerable literature on performing MR of DNAm on complex traits, and generally EWAS is performed using a model that allows for a causal interpretation and by testing the hypothesis that DNA methylation differences are driven by cellular reprogramming35. We believe this is the first opportunity that we as a field have had to examine the assumptions of MR in the analysis of DNAm on complex traits, due to the high yield of mQTL and the frequent
occurrence of multiple independent mQTL per probe. While some (most likely in the epigenomics or functional genomics fields) might have assumed the result, others (most likely in the epidemiology / common disease fields, where much of the 450k DNAm analysis is done) assumed the converse, and we are simply trying to present evidence that speaks to the question.
>>> Ok

In regard to common and non-haematologically driven diseases cannot the response statement that "These DNAm sites are not generally on the causal pathway mediating genetic effects on disease" be stronger and state no causal evidence? What are the generalisable implications of the weakness of the causal results (line 561) where only just over half were in the correct direction? What do the authors consider this implies when this methodology is performed for single traits – and again when considering the potential number of mQTL in the genome.
• The method that we advance here is to look for a general relationship between the observed causal effects from mQTL across multiple probes, and the expected relationship under a model of causality. The power of the study is to enable this where previously it was not possible. But the limitation is that it only speaks to the general relationship and not to the relationship at any specific probe. The point, however, is that often it is the mQTL instrumentation design itself that is used to make a claim about probe-specific causal relationships, and so the design of this analysis helps to indicate the limitations of that approach without an appropriately large number of independent instruments for a particular probe.
>>> These limitations discussed should be included the manuscript

Clear effects of traits influencing DNA methylation in blood, such as Dekkers et al triglyceride DNAm associations, are supported mechanisms (line 619). As the causal model is not supported (line 570), it would be highly useful for the common disease field if the authors can use the power of this study to state whether the conclusion of previous EWAS papers cited here, like Wahl et al. – which identified predictive findings in pathogenically non-relevant tissues as well as consistent findings across multiple cells, are robust? Or are these findings only biomarkers of trait-related changes and genetic confounding?
• We do conclude, having investigated several large scale EWAS studies, that "These results indicate that traits causally influencing DNAm levels in blood is the most likely mechanism that gives rise to these EWAS hits." (end of section on "Influence of traits on DNAm"). This result is echoed in the Implications. "Overall our data and results have resulted in the most comprehensive atlas of genetic effects to date. We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."

In the broader context, Lappalainen & Greally35 suggest that EWAS have been carried out to find loci that indicate cellular reprogramming. We found in this paper and others3,36, meaningful interpretation of EWAS findings is hampered by confounding factors and reverse causation. We agree that many EWAS studies do not account for genetic confounding or test for reverse causation. If we as a field are to find loci that indicate cellular reprogramming, then we should in addition to genetic confounding also take into account other confounding variables such as transcriptome and generate new reference panels to adjust for cellular heterogeneity. The need for second generation EWAS is now apparent in the EWAS field and we agree that our mQTL catalogue would be part of this as we anticipate that some previously reported EWAS associations are likely due to reverse causation. As EWAS currently stands, at best it uses DNAm sites as a marker for relevant regulatory functions, without being able to speak to function.
To address the issue of genetic confounding in EWAS large meta-analyses are needed and we will investigate this in a separate manuscript. With regards to reverse caution, Wahl et al.16 did show that DNAm can be altered in response to increased BMI. Dekkers et al.17 also showed that DNAm was changed in response to altered lipid profiles.
>>> The authors state above "to address the issue of genetic confounding in EWAS large meta-analyses are needed". Is this not what this publication is? – and exploring these genetic effects is a core justification for this publication. Yes, both Dekkers et al. and Wahl et al. show the physiological impact of high lipids on the blood methylome, directly and indirectly, respectively. However, Wahl et al. controversially implied that a CpG's DNA methylation change in peripheral blood was 'causal' in obesity– not just the signature of the phenotype-related hyperlipidaemia, hyperglycaemia, inflammation, etc. as others have found. Furthermore, Wahl et al. identified consistent DNA methylation changes across multiple cell types, which is not coherent with our knowledge that biological changes with the haematological system impact differently on particular blood cell-types [You et al. Nat Comms 2020]. The authors should use this large meta-analysis to help advance the field by being more definitive about previous results where they can.

The Selective analysis does not appear to be adding any substantial findings - as the conclusion given is only that a minority of regions containing DNAm sites have experienced positive selection and, significantly, no evidence is presented that the DNA methylation itself is driving any of the identified selection. This again also needs to be put in context that these data are only coming from 1.5% of the DNA methylome. Many additional SNPs not currently classified as mQTL would be if more CpG DNAm data was available.
• When performing the analysis to see if the mQTL were enriched for selection signatures, the question that we are trying to address is whether these variants, when compared against a background of variants with similar genomic properties, have higher overlap with genomic regions with selection signatures than expected by chance. Generally, an enrichment analysis does not have the complete set of causal variants available to its disposal (in a standard single trait GWAS, not all causal variants are known due to power), and the same is true here – we do not have all causal variants for DNAm. Our analysis reliably concludes that there are more selection signatures colocalising with our mQTL than we'd expect by chance, and that it is the fact that the positions associate with DNAm that drives the association (rather than other properties such as gene proximity, MAF, LD, etc). We feel that this is a reasonable and widely
adopted approach to examining a set of discovery loci and shouldn't be disregarded just because don't have knowledge of every mQTL. The results are still relevant to the mQTL that we do have – and that is a vastly larger number than has been recorded previously.
>>> Yes, the point that was raised wasn't to disregard the findings but for the authors to extend what the implications of their findings could mean.

As the authors state, the DNA sites are likely the primary target of selection - so the DNA methylation is just co-varying with genetic variation. In 19/42 at least one cis mQTL SNP with extreme SDS (Line 673) - but what would this be by chance with so many unidentified mQTLs? Can authors really say "these loci have been modified by disease pressure" (line 680) - with all the missing data? Genetic selection has shaped the genome and due to the sheer number of mQTL will be represented in the observed DNA methylome even if the majority of DNAm is not functional change. How does this impact on their conclusions for height and cardiovascular disease? (line 683).
• We have tested the mQTL for enrichment of selection and so we know that our signal is not purely by chance. We agree with the reviewer that "the majority of DNAm is not functional change", though note that the case here in still unclear as we are considering the set of loci with the strongest selection signals. We agree that the interpretation of why these mQTL are enriched for selection signatures is hard, and we changed the text to clarify the difficulty in assigning mechanism to the selection signals that exist at these loci. Specifically, whilst it is the most likely explanation, we indeed don't know "these loci have been modified by
disease pressure" and removed this claim.
We changed the text to: "A comparison showed that the genetic variance for cardiovascular disease associated mQTL or height associated mQTL with extreme SDS was higher when compared to all trait associated SNPs (Figure S49). To summarize, our results provide the first evidence that selection may have shaped the landscape of DNAm values across the genome although the mechanism for the selection signals that exist at these loci remains unknown."
>>> Good to acknowledge

Regarding the Intermediate methylation values, the authors should acknowledge that the two isolated white-blood-cell subsets of T cells (CD4+) and Monocytes (CD14+) are not homogeneous entities and that further sub-types of CD4+ and CD14+ are clearly present and observable epigenetically. Therefore, cell-to-cell variability (line 358) cannot be definitively concluded and cell-type differences may still be present. These intermediate loci could, in fact, fit with loci that are enriched for important functional units that are specifically driving or represent these divergences into further subcell-types.
• We thank the reviewer for this comment and changed this sentence to: "By replicating this finding in two isolated white-blood-cell subsets(Figure S22), we showed that this is due to cell-to-cell variability16,37 or sub cell type differences which may indicate that these loci contribute to the divergence into further sub cell types."
>>> Good to acknowledge

In regard to the Authors reply re results "support a mixed genetic architecture of small genome-wide polygenic effects and larger cis effects." The statement that "relatively little left to be discovered in cis" – it is unclear how the authors can state this if only assessing 1.5% of the DNA methylome and genetic variation is currently only limited to variants well-tagged by common SNPs (much structural variation, repetitive genome, and poorly sequenced regions of the genome remains). ~50% of the all CpGs in the human genome reside within the Repetitive genome that is very poorly covered by arrays for technical reasons.
• Again, we apologise that the wording in the text may have strayed from emphasising that the 450k array only represents a small proportion of the methylome. We have changed the text to clarify that we only assessed DNAm sites on the 450k array: "We then investigated how much of the heritability of variable DNAm can be explained by our mQTL associations on the 450k array using family-based heritability studies of DNAm33,38"
>>> Good to acknowledge

Furthermore, is not stating that the genetic architecture of DNAm is 'polygenic' rather obvious, as no previous experimental or genome-wide evidence implied it was monogenic or would fit with its epigenetic role? This meta-analysis points to a likelihood that mQLT will exist for almost all CpGs, except those few that are squarely within high dense CGI that are almost always unmethylated.
• Although this seems obvious no other study had the power to conclude this or focused on cis mQTL only or trans mQTL of GWA SNPs.
>>> Ok

Further to this, the authors state the "very small genetic effects on DNAm is valuable for gaining power to evaluate whether its variation has a substantial causal role in disease and other biological processes". Conceptually the authors have not commented on the implications of their findings if 45% of array probes are potentially influenced genetically by mQLTs. Even extrapolating at this level implies potentially 12 million CpGs, but additionally, they estimate that with each probe there was an increase of 0.76 cis mQTL and 0.05 trans mQTL – which could push this up to a level of ~21 million CpGs being genetically influenced. This fits with the fact that this 450k array is more biased towards those CpGs nested within very high CpG dense CpG islands regions – which show extreme lack of variability (unmethylated on any allelic background - as confirmed in Figure 1b) so that the level of genetic influence of even ~45% is an underestimate. Therefore, it is clearly very easy to find mQTL even in this meta-analysis of these reasonably homogenous population groups.
• We agree that many probes had low variability. The implication here is that a genetic effect will be harder to detect because the signal to noise ratio in the probe assay is lower, so the statistical power is lower. However, tranches of probes that have higher variability are not guaranteed to have more easily detectable genetic effects – they could be less genetic and more stochastic or unregulated. A small whole methylome sequencing study of twins5 demonstrated
that only a small proportion (~15–20% of 28M DNAm sites of the methylome) is differentially methylated in the population and that non-shared unique environmental factors are the main source of variation. Extrapolating the number of expected associations will depend on variance component analysis of probes that we do not have assayed (as DNAm sites will have variable genetic and environmental contributions) and is a worthy future study.
We moved the Implications paragraph about coverage to the Results:
"The microarray technology used in the majority of cohorts limited us to analyse only 2% of sites across the genome12, which are biased to promoters and strongly underrepresented regulatory elements. To explore the impact of expanding the coverage of arrays, we calculated the linear relationship between the median number of probes by gene on the 450k array and the median number of cis and trans mQTL. For each probe, we found an increase in the median of 0.76 cis mQTL ($p<9.03e-16$) and 0.05 trans mQTL ($p<1.47e-05$) (Figure S9). A similar increase was seen in non-genic regions. This indicates that expanding coverage will increase mQTL yield although this will depend on the genetic contribution of the DNAm site."
>>> Ok – but would be good to state how many total CpGs could have mQTL even with the inadequate assumption of a similar rate. As discussed above, the authors should focus on what clear conclusions can be draw from their large array meta-analysis – not require support from less robust conclusions from an individual small low-coverage WGBS experiment.

Additionally, future increases in power may bring more nuanced and further independent signals from their current LD clumped cis effects (without definitive knowledge that all effects can be completely coalesced down to only 1 variant). What are the implications regarding the inferences of causality if the vast majority of the DNA methylome has genetic influence? What does that mean in terms of the

interpretation of MR with methylation?
• We added the following sentences in the Implications to discuss this:
"Fourth, DNAm might be on the causal pathway in a disease-relevant cell type or context."
..
"Finally, EPIC arrays2 and promising low-cost sequencing technologies19 offer an opportunity to make detailed interrogations of enhancer and other regulatory regions, which may be more fruitful in finding causal relationships with complex disease."
>>> Therefore, the authors need to include directly a comment on caution due to the potential for over-interpretation.

How could this contribute to the development of second-generation EWAS as they state? (Line 730).
• We agree with the reviewer that this should be clarified. We changed this sentence to: "We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."
>>> Ok

Also, what is the potential of this increased power to detect more subtle technical variation - as well as direct genetic effects identified? Therefore, as the potential cis effect in the DNA methylome is extremely large, how can the authors definitely state that "Yet the majority of genetic effects are likely to act in trans?"
• We agree with the reviewer that these are important questions and they do naturally stem from the results that we have presented. We hope that our study can help shift the conversation in that direction, and that it is clear that the field as a whole needs to endeavour to explore these questions, and given how packed the paper is already it is difficult to extend the scope further."
>>> The authors need to more clearly state what they mean by this – re effect size of cis versus trans and how the current evidence implies the majority of genetic effects act in trans – also taking into consideration the relative genome space size of cis versus trans.

The authors consider that this meta-analysis approach brings unique insights through increased power. Therefore, it would be good to convey more strongly how the synthesis of these data could drive the field forward, rather than supporting conclusions from older papers - many of which will now have been superseded due to recognised potential technical issues.
• Thank you for making this point. We feel that an important conclusion of this paper is not being assimilated – additional power to detect genetic effects has led to us attempting to understand how perturbation of the methylome leads to disease and functional variation, and we have found little in terms of these relationships. The major biological finding is that variation in the methylation sites we assayed does not have obvious functional effects for disease and either the genotype-phenotype map is far more complex than we currently understand, or DNAm at these sites in bulk tissue should be de-prioritised in understanding disease variation. Whether DNAm generally is on the causal molecular pathway to disease, when interrogated in a cell-type specific manner and having accounted for wide ranging pleiotropic possibilities, is not known.
We have now updated the Implications and added:
"We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."
>>> Yes, this is in fact the point in terms of driving the field forward - that as the author state, there are "not obvious functional effects for disease". Therefore, the request was to interrogate previous results and determine if their potential claims of functionality are incorrect.

Also, it would be good to try to integrate the latest functional insights to the DNA methylome such as the site-specific tendencies of site-specific dependencies of DNMT and TET activity (Ginno et al Nature Comms 2020) and specific repressive and active roles with specific transcriptions factors (Yin et al Science 2017, Domcke et al Nature 2015) etc.
• We now introduced the TF section as follows:
"Recent studies have uncovered multiple types of transcription factor (TFs)/DNA interactions with DNAm including the binding of DNAm-sensitive TFs24-26. Epigenetic editing studies have revealed that local methylation and demethylation activities are affected by TF binding and cooperativity between TFs25,27."
>>> Good to acknowledge

Additional Major Points
1. "The role of inter-individual variation in DNA methylation on disease mechanisms is not yet well characterised" – Incorrect statement, the authors are referring to 'common' variation only and should clearly specify this – robust pathogenetic involvement of DNAm in rare disease such as Fragile X, Facioscapulohumeral muscular dystrophy, MLH1-deficient sporadic colorectal cancers, etc.
• We changed the text to: "The role of common inter-individual variation in DNA methylation (DNAm) on disease mechanisms is not yet well characterised."
>>> Good to clarify

2. The authors only cite papers re genetic influence on DNA methylation from 2014 onwards – but exclude landmark earlier work from Kerkel et al. Nature Genetics 2008 and others. Furthermore, the McRae et al. citation is problematic as they confused genetically-driven DNA methylation differences with the distinct biological mechanism of "transgenerational" epigenetic effects (only robustly seen in plants - Horsthemke N Comms 2018) – by including "transgenerational" in their title.
• Thanks for these comments. We have removed the McRae et al. citation from the following sentence "Because genetic influences on DNAm in blood have been shown to be widespread33,34". We do use BSGS heritability estimates later on from the McRae paper38 which we don't feel are controversial. Kerkel et al.43 (which is indeed a landmark study) describes allele specific methylation which is a comparison of alleles in a heterozygous individual whereas we refer to the association between genetic variants and DNA methylation in multiple individuals.
>>> It would be preferable to use other heritability estimates and not to cite and further propagate in the literature this erroneously titled paper in what will be a high-profile consortium paper. This fundamental error regarding 'transgenerational' has already caused significant confusion in the field. The Kerkel et al. paper was the first to show in humans the relationship between SNPs and methylation, by allelic variation, which is one of the drivers of the population variation observed.

3. A conditional analysis within the locus of each mQTL identified 758,130 putative independent variants (line 295). Again, this needs to be put into context that only 1.5% of the DNA methylome is being measured, that median 2 and IQR 4 independent variants of just these CpGs and therefore how large the potential genetic to methylation space is.
• We moved the paragraph describing the limitation of the array from the Implications to the results to clarify this.
"The microarray technology used in the majority of cohorts limited us to analyse only 2% of sites across the genome12, which are biased to promoters and strongly underrepresented regulatory elements. To explore the impact of expanding the coverage of arrays, we calculated the linear relationship between the median number of probes by gene on the 450k array and the median

number of cis and trans mQTL. For each probe, we found an increase in the median of 0.76 cis mQTL (p<9.03e-16) and 0.05 trans mQTL (p<1.47e-05) (Figure S9). A similar increase was seen in non-genic regions. This indicates that expanding coverage will increase mQTL yield although this will depend on the genetic contribution of the DNAm site."
>>> Good to acknowledge

4. The authors state that the cohorts have diverse geographical origins (line 689) – but these are only European ancestry population cohorts. What implication do their findings have regarding potential genetic effects in studies performed across population ancestry groups or with mixed population groups?
• We think it is vitally important to include non-European studies. Given the technical challenges of this study we wanted to restrict to a single ethnic group to avoid potential confounding that would be difficult to guard against. With this study as a reference we are now well positioned to analyse multi-ethnic studies and going forward this will be a major priority.
>>> Therefore 'diverse geographical origins' is not correct

5. Are these results impacted on by the identification of further 450k array probe technical variability issues as described by Sugden et al. Patterns 2020?
• We used two approaches to exclude DNAm sites from our analyses. First, we excluded 50,186 DNAm sites that were masked by Zhou et al.44 which includes probes with potential cross reaction and probes that could not be mapped to genome. Secondly, we removed an additional 14,882 probes including multimapping probes (bisulfite converted sequences allowing two mismatches at any position mapped to the hg19 primary assembly) and probes with variants (MAF >5%, UK10K) at the CpG dinucleotide or the extension base (for type I probes).
Overall, we found that the reliability estimates from Sugden et al.45 were slightly lower for probes that we analysed (median = 0.196 vs median = 0.082). We then compared the reliability estimates to the five categories described in Zhou et al.44 and found that probes that were affected by a SNP have high reliability estimates (median= 0.502 vs median = 0.09). We also extracted reliability estimates for our mQTL probes and found that the reliability was much higher (median= 0.266 vs median = 0.0356) indicating that the type II error was small (which makes a lot of sense, since you can't have replication between cohorts if you can't measure the probe reliably). However, we may have missed genetic effects on unreliable probes.
>>> Ok, good to incorporate this new analysis

6. Statements such as "DNAm is usually associated with gene repression, with loss of DNAm reflecting enhancer or gene activation" (line 339) are an oversimplification and need to be clearly caveated in regard to the promoter focus of this array and the authors' interpretations – i.e. ignores important role in gene body methylation transcription and integration of the epigenome with chromatin modifications (Baubec et al Nature 2015 etc.), identification of TF that require DNA methylation for binding (Yin et al Science 2017) etc.
• We have updated the wording with findings from these papers: "Recent studies have uncovered multiple types of transcription factor (TFs)/DNA interactions with DNAm including the binding of DNAm-sensitive TFs24-26. Epigenetic editing studies have revealed that local methylation and demethylation activities are affected by TF binding and cooperativity between TFs25,27."
>>> Good to acknowledge

7. Authors need to state where findings are as expected and have potential clear explanations – e.g. 'Sites that overlap TFBS are relatively hypomethylated'. TBFS are enriched within high CpG density

loci (i.e. CpG islands) that are generally always unmethylated and 'TFBS enrichments were not tissue specific' also fits with these residing within CpG islands – which clearly overlap with promoters which are less tissue-specific.
• We have highlighted expected findings in the text:
"However, we found that the highest correlations were for associations involving trans-only sites (Adipose rb=0.92 (se =0.004); Brain rb=0.88 (se=0.009)) despite having on average smaller effect sizes than cis only associations, implying that they are less tissue specific than cis effects (Adipose rb=0.73 (se =0.002); Brain rb=0.59 (se=0.004)) which is line with the notion that promoters are less tissue-specific."
…
"We found strong enrichments for the majority of TFs amongst DNAm sites with a trans association (cis+trans: 55%; trans only: 80%; cis only: 18%) which is in line with the observation that loss of DNAm at promoters is usually associated with gene activation30, and amongst cis-acting SNPs (cis only: 96%, cis+trans: 91%, trans only: 1%) (Figures 2B, S28, S29). Consistent with the observation that trans only DNAm sites are enriched for CpG islands (Figure S19), sites that overlap TFBS were relatively hypomethylated independent of tissue (weighted mean DNAm levels = 21% vs 52%, p<2.2e-16) (Figure S30) and we found that generally the TFBS enrichments were not tissue specific (Table S10-11, Figure S28-29)."
>>> Good to acknowledge this known biology

8. The authors cite an old paper for support that "trans-SNPs were typically found in intergenic regions which are depleted for complex trait loci" from Schork et al. PLoS Genetic 2013. Are this paper's conclusions consistent with current knowledge of predominately noncoding GWAS findings - as was summarising early arrays used for GWAS from 2010/2011?
• We agree with the reviewer. Approximately 90% of the phenotype associated SNPs identified by GWAS lie within non-coding regions. After accounting for LD, enrichment analyses have shown that GWA SNPs are enriched for coding, conserved and regulatory regions such as promoter and enhancer regions 46(Finucane et al. 2015). We didn't find any enrichments for these regulatory annotations for trans only SNPs.
"These enrichments correspond to the results reported earlier, in which trans-SNPs were typically depleted for enhancer and promoter regions, whereas complex trait loci are enriched for coding and regulatory regions47."
>>> Ok, good to clarify

9. "1,373 putative examples in which at least one DNAm site putatively shared a genetic factor with 71 traits including only 19 disease (line 540). Again, how do authors interpret this if genetic influences on DNAm are ubiquitous?
• This doesn't seem to be a contradiction – not every variant will influence DNA methylation, and not every variant will influence complex traits, and the overlap doesn't need to be complete?
>>> The comment was not regarding this being a contradiction and the requirement for complete overlap – but with so many SNPs already impacting on observed CpGs and the implications of all the unassayed CpGs - and what this means for the potential over-interpretation of current positive findings.

10. Regarding the GWAS signal and DNAm site colocalised analysis (line 545). What is the true relevance of this closest determination – as the authors don't have exhaustive CpGs data in the region so may have no data other close CpGs and, furthermore, the GWAS signal unlikely to be the actual causal functional SNP in each locus?

• Coloc doesn't require knowledge of the causal variant, it just looks for consistent pattern of association at a locus across two traits. We likely don't detect every instance of a shared causal variant, but we only make inference about the ones that we are able to.
>>> However, the authors need to acknowledge the large number of missing CpG data in this array analysis.

11. For the widespread changes in DNAm levels (line 599) can the authors give an estimate of how many CpGs are potentially contributing for each these 5 outlier traits?
• The manner in which this was calculated does not allow us to count the precise number of CpGs that are influenced by these traits, as we are comparing the distribution of observed test statistics against the theoretical distribution under the null hypothesis of no associations.
>>> Ok

Reviewer #3:
Remarks to the Author:
The authors have addressed all my comments. I have no additional comments.

Author Rebuttal to Initial comments

Referee expertise:

Referee #1: QTLs, computational biology

Referee #2: epigenomics

Referee #3: QTLs, omics data

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
I believe that the authors have addressed all of the major concerns from my initial review. I really like the simulation study that you have performed for the colocalisation analysis. It address the issues that I had about potential false positives in the analysis.

Minor comment: On line 681-682 of the supplementary data you write that "… we use the sparse approach for colocalisation analysis when **incomplete** regional data around the mQTL is **unavailable**." This double negative is a bit hard to parse, did you mean complete instead of incomplete?

Reply: Thanks for this. We changed:

"In light of these simulations, throughout this work we use the sparse approach for colocalization analysis when incomplete regional data around the mQTL is unavailable"

to

"In light of these simulations, throughout this work we use the sparse approach for colocalization analysis when complete regional data around the mQTL is unavailable".

I also appreciate that you have added the caveat to the discussion that "Fourth, DNAm might be on the causal pathway in a disease-relevant cell type or context.". This addresses my concerns about potential cell type specific effects. I agree that it is currently not feasible to systematically evaluate this possibility due to the lack of appropriate cell-type-specific data. I also agree that your result about meQTLs in whole blood not generally being on the causal path for complex traits is still a very important finding, especially in the context of EWAS studies that are often based on the same tissue and DNA methylation array. I would definitely recommend all future EWAS studies to very carefully consider your results. I also believe that many of the same issues might be encountered when trying to use bulk tissue trans-eQTLs in causal inference.

Reply: Thanks for this comment. In response to comments from Reviewer 2 we emphasized this further:

"While our results only include <2% of DNAm sites in the human genome and are limited by the two-phase design, these findings have several implications especially in the context of EWAS studies that are often based on the same tissue and DNAm array."

and

"However, as disease relevant signals and regulatory regions may be cell type specific, new analytical tools are required to infer cell type specific mQTL from bulk tissue."

Below are two minor comments:

I have now tried to read the two-dimensional enrichment analysis multiple times, but I am still not sure I understand how to interpret the results. Let's look at the IRF1 example. As I understand it right now, you have found that DNA methylation sites near the binding sites of the IRF1 TF are enriched to have trans-meQTL SNPs near the binding sites of EZH2, SMC3, ATF3, BCL3, TR4 and MAX TFs. Since you also mention that these trans-meQTLs could be driven by TF expression (line 467), does this imply that the binding of EZH2, SMC3, ATF3, BCL3, TR4 and MAX factors at these distal sites somehow influences the expression (or activity) IRF1 via some uncharacterised trans-acting factors? And then the change in IRF1 expression is what ultimately influences the DNA methylation of the target sites? I believe the analysis is sound so I don't think this issue should hold up the publication of this study, I was just struggling to understand what the interpretation of the results would be.

Reply: Our interpretation in the paper is as follows:" To summarise, our results show that *trans* mQTL are in part driven by long-range cooperative TF interactions". We speculate that our findings are in line with the TF indirect collaboration model suggested by Domcke et al. (Nature 2015). This model proposes that binding of DNA-methylation-sensitive TFs relies on additional determinants to induce local hypomethylation. Domcke et al. found that NRF1 binding *in vivo* critically relies on the local DNA sequence context in *cis* and TFs in *trans* to ensure a hypomethylated binding site.

We changed the following sentence from:

"We hypothesized that if a *trans* mQTL is driven by TF expression[7,9] then particular TF-TF pairs may exhibit preferential enrichment[31]."

into

"We hypothesized that if a *trans* mQTL is driven by TF activity[7,9] then particular TF-TF pairs may exhibit preferential enrichment[31]."

I am still not convinced that it is meaningful to split trans-meQTLs into cis+trans (covering also cis-close-to-trans) and trans-only categories, but since this distinction does not significantly change the major conclusions of the manuscript, I accept the authors decision to keep it as it is. However, there's one more thing that I would like the authors to check: are the trans-only meQTLs significantly further away from 450k probes than cis+trans and cis variants? My reasoning is that maybe trans-only meQTLs do not have any cis associations just because they happen to be far away from the sparse set of DNA methylation sites profiled by 450k arrays (i.e. they are near distal regulatory elements not captured by 450k arrays). Thus, perhaps you would also find cis associations for these if your were to profile the DNA methylation using a technology that has higher coverage of the genome. This would be easy to check by comparing the distribution of distances from the lead meQTL
variant to the nearest 450k probe included in your analysis for the three meQTL categories.

Reply: There is indeed a difference between *trans only* SNPs, *cis+trans* SNPs and *cis only* SNPs and the distance to the closest 450k CpG. However, the same could be true for *cis only* SNPs as we may not have found the *trans* mQTL due to the two-phase design and/or power. The results of these enrichments are patterned, but clearly only relate to the ascertainment properties of 450k array and the study design.

Reviewer #2:
Remarks to the Author:
Responses to Min et al. replies to
Referee #5 (Remarks to the Author):
>>> Indicates Reviewer Reply

Min et al. have performed a multiple European ancestry cohort mQTL meta-analysis (27k with 5k replication) from 36 published cohorts in DNA derived from peripheral and cord blood. This analysis was performed with the previous generation Illumina DNA methylation 450k array. ~10M genotyped SNV were used (imputed 1000G panel). From 420k post-QC CpG probes they identify that ~45% (190k) of all the DNA methylation sites on the array are influenced genetically by identified mQTL – totalling >270k independent mQTL. Due to the power of this study, this was almost a 2x increase in cis- and 10x in trans associations, respectively, compared to a similar study on one of these cohorts by Bonder et al. This later trans finding included 23k independent mQTLs (8.5% of total). This expansion in trans findings - with the high statistical threshold ($p<1e-14$) required for these due to the large testing space - is a highlighted strength of the meta-analysis. These mQTL in total are seen to explain
15-17% of the additive genetic variance of DNAm. Well-acknowledged statements are made regarding the genetic influence on DNAm levels being highly polygenic that clearly fit with previous experimental and genome-wide analyses.
Whilst shared genetic factors associated with both blood DNAm levels and complex diseases are identified, no strong casual relationships from DNAm to trait are supported – again in line with many current common DNAm EWAS analysis interpretations. The analysis of the expanded trans results are further extended into networks, although without any functional evaluation to further support these. Selective findings state that DNAm exhibits signatures of negative and positive natural selection, but do not provide evidence that the epigenetic change is itself causally driving this. The use of the now superseded 450k array clearly enables large numbers of samples from multiple cohorts to be included in this metanalysis. >90% of these probes are still included on the later EPIC array. Also, it is still useful for generalisable results as long as the bias of this array is acknowledged. Additionally, as the majority of EWAS to date have been performed with this 450k platform, enabling
further reinterpretation of these previous studies' conclusions which could be very beneficial.
Similar work has been performed in some of these cohorts individually before and this extends on papers as mentioned from Bonder et al. Nature Genetics 2017, as well as Blueprint consortium Chen et al Cell 2016, etc. Whilst, older papers can be used to justify points made, instead, it would be good to see that these data can convincingly lay down the state of current knowledge. Referring to DNAm as a complex trait does make for some chunky language at times and can lead the reader to inflate the potential functional significance of every detected mQTL.
• After careful checking we now changed two sentences that refer to DNAm as complex traits. We changed
"As with classical complex traits31, much larger sample sizes are required to map associations involving small genetic effects in order to permit greater understanding of the genetic architecture and the biological processes underlying DNAm32."
to
"Much larger sample sizes are required to map associations involving small genetic effects in order to permit greater understanding of the genetic architecture and the biological processes underlying DNAm32."
and
"This revealed a genetic architecture of DNAm that is polygenic, as with many traits traditionally considered 'complex'."

to
“This revealed a genetic architecture of DNAm that is polygenic.”

>>> Good to acknowledge and improve this

A recurring issue detailed in various points below is that the authors need to keep in the forefront of all of their conclusions that they are only testing 1.5% (420k of 28M CpGs) of the DNA methylome. Whilst statistically some short-range co-methylation can be detected, the limited array findings do not benefit anywhere near to the same level as genetic arrays do via LD in capturing common variation. The implications of this from the current findings – causality, variance, heritability, cis+trans identification, selection analyses etc need to be acknowledged.
• We apologise that the wording in the text may have strayed from emphasising that the 450k array only represents a small proportion of the methylome. This is something that we are keenly aware of and highlight the point in the Implications section at the end of the paper. However, as the reviewer rightly points out, these are the sites that have been under analysis by the field, and our paper only seeks to evaluate the relevance of these sites. The extent to which this speaks to the total relevance of DNAm genome-wide remains to be seen, but it must be noted this paper identifies the largest number mQTL ever discovered on a much larger number of participants and studies than in previous efforts. The analysed sites represent 99% of RefSeq genes with multiple probes per gene, 96% of CpG islands from the UCSC database, CpG island shores and shelves, FANTOM4 promoters, the MHC region and some enhancer regions and additional content1 (so, to be fair they are more like extended exome
arrays than genome-wide arrays if a comparison to SNP chips must be made). Although, underrepresented in distal regulatory elements (i.e. enhancers) and representing only 1.5% of all DNAm sites across the genome, these sites are actually quite representative of the known functional methylome. We agree that extrapolating our findings beyond the measured sites is not possible in this study. Rather, the purpose of this study is to evaluate the properties of the sites that have been routinely analysed in hundred thousands of datasets for the past decade (101237 datasets uploaded in Gene Expression Omnibus as “GPL13534”), in EWAS studies and for reference genomes including BLUEPRINT3, International Human Epigenome Consortium4 and The Cancer Genome Atlas consortium (https://tcga-data.nci.nih.gov/tcga/). We would like to acknowledge that profiling the entire methylome would also not be sufficient to explain the underlying complex genotype-methylome map as the CpGs will have variable genetic and environmental contributions. A whole methylome sequencing study of twins5 demonstrated that only a small proportion (~15–20% of 28M CpGs of the methylome) is differentially methylated in the population and that non-shared unique environmental factors are the main source of variation.

>>> As stated, the major and useful benefit of this current consortium analysis is to be able to re-examine the large number of previous studies that have been performed with this now superseded array platform. However, the additional statements by the authors that this 450k array captures an equivalent portion of epigenomic functional information as the almost complete coding genetic information gathered from an “extended exome array” conveys a poor understanding of the epigenome. This array DNA methylation array is not, as the authors state, “actually quite representative of the known functional methylome” – it is was designed a decade ago and was focused on promoters and genic regions. Beyond the fact that this array only covers only 1.5% of the total CpGs, all of our accumulated knowledge of the DNA methylome since this time (ENCODE, Roadmap etc) has identified that promoter regions, being CpG dense, are highly invariant. They remain hypomethylated across all tissue

types, making many of these CpGs on the array redundant. Furthermore, genic regions are not found be to be enriched for other regulatory information, instead this predominately resides beyond in distal non-coding regions. The array does include coverage of a representation of CpGs that reside in more dynamic intermediate density CpG shore regions, but lacks any of the focused coverage of currently known enhancer regions included in its successor, the EPIC 850k array (EPIC itself still only covers 58 % of FANTOM5 enhancers, and 7 % distal and 27 % proximal ENCODE regulatory elements [Pidsley et al., Genome Biol. 2016]). These enhancers regions are considerably more dynamic and significant in term of driving functional lineage changes etc [Heintzman at al., Nature 2009, Ziller et al, Nature 2013, etc]. Additionally, all current methodologies poorly represent the ~50% of CpGs that reside within repetitive genome. Increasingly accumulating evidence indicate that these loci play a significant pathogenic role in disease-related regulation [Chuong et al., Nat Rev Genet. 2016] by perturbing transcriptional regulatory networks. The authors need to be more circumspect about what they are actually assaying and its limitations.

Reply: We apologise for using the term 'known functional methylome' in the reply, though we note that it is not used in the text and is not implied to be the case either. We do appreciate that we are only analysing less than 2% of DNAm sites in the human genome and the revisions emphasize this, and we do not extrapolate beyond that in the text. The main text of the manuscript does describe the content of the 450k array: "The microarray technology used in the majority of cohorts limited us to analyse <2% of sites across the genome[12], which are biased to promoters and strongly underrepresented regulatory elements."

We now emphasized this further in the Implications section:

"While our results include <2% of the DNAm sites in the genome and are limited by the two-phase design, these findings have several implications especially in the context of EWAS studies that are often based on the same tissue and DNAm array."

and

"Future studies may be more fruitful in finding causal relationships with complex disease. Either EPIC arrays[13] or low-cost sequencing technologies[69] will expedite detailed interrogations of enhancer and other regulatory regions. Especially, single molecule long-read sequencing promises to expand the genetic and epigenetic spectrum by allowing the detection of complex genetic variation such as allele specific DNAm and structural variation and different types of DNA modifications. However, as disease relevant signals and regulatory regions may be cell type specific, new analytical tools are required to infer cell type specific mQTL from bulk tissue. Given our projection of mQTL yields expected for future studies, pleiotropy involving mQTL is likely to be increasingly important to model when interpreting genotype-trait pathways."

>>> It is unclear what point to help justify their meta-analysis array study that the authors are citing Busche et al. - a low coverage whole genome bisulphite methylome sequencing study of 30 twins - for? Clearly looking at a larger fraction of the DNA methylome would identify more genetic effects. Firstly, even that study identified ~15–20% CpGs were differentially methylated, which is in fact still ~4-6 million CpGs, so not a "small" as the authors state. Secondly, that study's coverage was <10x, therefore, significantly less than the Roadmap standard of 30x. Its conclusions are limited to the resolution that that level of coverage can convey. Detection of small percentage DNA methylation changes, which robust array studies have shown us are the predominate observations in the DNA methylome, are required to interrogate fully common factors; such as genetic influence and cell-type heterogeneity. Furthermore, saturation analysis

has shown that in fact ~100x is required for robustly differential methylation analysis at individual CpGs [Libertini et al., Nat Biotech. 2016]. Without robust high-resolution data, delineating the proportion of genetic to non-genetic effects cannot be concluded. In fact, more accurate array studies in monozygotic twins have found the reverse, by removing the large genetic component leading to almost a complete loss of identifiable disease-related DMPs compared to population studies [Paul et al., Nature Comms 2016]. It is the robustness of the CpGs probed, although limited in number, by these Illumina DNA methylation arrays that is the clear advantage of this methodology over low-coverage bisulphite sequencing. Subsequently, conclusions may be able to be drawn, albeit for these small number of CpGs, with potential broader implications for the complex genotype-methylome map.

Reply: We remain in agreement with the reviewer and have no argument against the notion that there is likely substantially more variation in the methylome than is assayed by the 450k array. We do believe that the main text is faithful to the notion that inference cannot be extrapolated beyond these features of the methylome. The only place where we extrapolate beyond the 450k sites is where we make a tentative prediction of the expected increase in mQTL that would be found as DNAm assay density increases (see text below) which we believe serves to enforce the general point that there is further variation that is beyond the 450k assay.

"The microarray technology used in the majority of cohorts limited us to analyse <2% of sites across the genome[12], which are biased to promoters and strongly underrepresented regulatory elements. To explore the impact of expanding the coverage of arrays, we calculated the linear relationship between the median number of probes by gene on the 450k array and the median number of *cis* and *trans* mQTL. For each probe, we found an increase of 0.76 *cis* mQTL ($p<9.03e-16$) and 0.05 *trans* mQTL ($p<1.47e-05$) (**Figure S9**). A similar increase was seen in non-genic regions. This indicates that expanding coverage will increase mQTL yield although this will depend on the genetic contribution of the DNAm site and cell type specificity."

• We have moved the analysis on the array and the mQTL coverage from the Implications to the Results to clarify this issue.
>>> Good to acknowledge and clarify this

Clearly the scale of this meta-analysis is a major undertaking, for which the authors should be commended in pulling together. It would, therefore, be good to further integrate the results with the latest functional understanding of the DNA methylome, where feasible, whilst clearly acknowledging any potential weaknesses due to lack of data from all of the DNA methylome. The authors state that they have identified a more complex genotype-phenotype map than has previously been hypothesised. These insights could be more strongly stated in the Abstract and Conclusion, if possible, particularly in regard to EWAS interpretation of previous studies. Further discussion of the responses to Reviewer 2 points as well as some additional comments are detailed below for the authors to consider.
• We thank the reviewer for this comment and we now changed the Discussion. We didn't change the abstract due to the word count.
"Fourth, DNAm might be on the causal pathway in a disease-relevant cell type or context.
…
Finally, EPIC arrays2 and promising low-cost sequencing technologies19 offer an opportunity to make detailed interrogations of enhancer and other regulatory regions, which may be more fruitful in finding causal relationships with complex disease."
…
"Overall our data and results have resulted in the most comprehensive atlas of genetic effects to

date. We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."
>>> Good to acknowledge and improve this

Points from Responses to Reviewer #2
In regard to definitions used here for cis and trans – consistency with previous publications (GTex etc) is a fair explanation and a good approach. However, the definition of trans does lead to a less clear and the more overly descriptive nature of the results with the potential dilution of 'true' trans effects by cis. So, the biological interpretation of the 80% of trans that are inter-chromosomal is clearly more straightforward. That "cis and trans mQTL operate through distinct mechanisms" (line 337) also fits with known experimental work from Stadler et al., Leinert et al., and others and this has been well conveyed by previous genome-wide publications, such as the Bonder et al. Paper.
• There is a distinction to be made between trans effects that may be biologically cis because they are on the same chromosome, versus trans variants that are proximal to other variants that influence DNAm in cis, or are themselves additionally influencing DNAm in cis. For the first case, we evaluate the extent to which omitting intra-chromosomal trans-effects are driving any enrichments and find that they are not. For the second case, we separate variants into categories that include 'trans-only' - in order to ensure that the trans-signals in enrichment analyses are not contaminated by potential cis-signals. This is something that previous studies have not done and has led them to make conclusions about trans operations that we can show are potentially unreliable.
>>> Fair enough

• We clarified this in the manuscript: "Our analysis shows that trans-only sites and SNPs have different properties as cis+trans SNPs and sites, indicating that enrichments of general trans categories are dominated by their cis functionality."
>>> Ok

It would be helpful if the authors could more clearly convey to the reader the mechanism of the cis effects observed and what this implies for their potential functionality. The DNA methylome is encoded via CpG Density and TFBS motifs – and cell-specific utilisation of this baseline state (as in the cited Schubeler review). Population allelic variation in both these genetic factors is a clear mechanism for the majority of cis effects described, as well as captured by variation in LD with these effects (Greally, Current Opinion Systems Biology 2017). This also encapsulates simple proximity CpG Density effects (Grandi et al. Genome Research 2015) and sequence-dependent Allele-Specific Methylation (Kerkel et al. Nature Genetics 2008). The genetic effects observed here were, as expected, identified to be highly stable across cohorts. As Figure S3a shows the mQTL and CpG couple, on the whole, reside within very close proximity of each other (This figure should potentially be included in the main paper to help convey this). The large impact of genetic effects on population
EWAS has been clear from comparison with Monozygotic Twin EWAS that led to almost complete loss of phenotype-related DNAm findings (e.g. Paul et al. Nature Comms 2016). So, the high level of genetic effects identified here on DNAm in this meta-analysis is not surprising. Chen et al. (Cell 2016) identified considerable genetic influences even in isolated immune cell epigenomes and specifically urged caution in EWAS interpretation because of this – which the authors could further explore as detailed below.
• We agree with the reviewer that other possible cis mechanism for genetic variation exists such as sequence-dependent allele specific DNA methylation or structural variation. Here, we focused on population allelic variation captured by genotype and DNAm arrays which is of interest for EWAS and other population-based studies. Future studies using long-read

sequencing where someone can determine both the DNAm level and the genotype from a single molecule could reveal biological insights of these mechanisms (allele specific methylation and structural variation).
>>> Including discussion of this beyond just reporting associations would be informative for those readers from the non-epigenetic community.

Reply: We changed the following text:

"Finally, EPIC arrays[12] and promising low-cost sequencing technologies[68] offer an opportunity to make detailed interrogations of enhancer and other regulatory regions, which may be more fruitful in finding causal relationships with complex disease."

to

"Future studies may be more fruitful in finding causal relationships with complex disease. Either EPIC arrays[13] or low-cost sequencing technologies[69] will expedite detailed interrogations of enhancer and other regulatory regions. Especially, single molecule long-read sequencing promises to expand the genetic and epigenetic spectrum by allowing the detection of complex genetic variation such as allele specific DNAm and structural variation and different types of DNA modifications. However, as disease relevant signals and regulatory regions may be cell type specific, new analytical tools are required to infer cell type specific mQTL from bulk tissue. Given our projection of mQTL yields expected for future studies, pleiotropy involving mQTL is likely to be increasingly important to model when interpreting genotype-trait pathways."

• In our paper we explain this mechanism in the Introduction: "Because genetic influences on DNAm in blood have been shown to be widespread33,34, a powerful avenue into researching the functional consequences of changes in DNAm levels is to map genetic differences associated with population-level variation, identifying DNA methylation quantitative trait loci, (mQTL) that include both local (cis mQTL) and distal (trans mQTL) effects."
>>> Ok

Causality and DNA methylation - DNAm sites have shared causal variants with complex traits via colocalization analysis - but the authors only have data on 1.5% of the DNA methylome. Therefore, the number of DNAm sites varying with casual variants is significantly larger. The authors state that "emerging ideas that DNAm variation is a marker of regulatory states and not a primary causal driver" – however, isn't this an acknowledged idea in the field?
• There is considerable literature on performing MR of DNAm on complex traits, and generally EWAS is performed using a model that allows for a causal interpretation and by testing the hypothesis that DNA methylation differences are driven by cellular reprogramming35. We believe this is the first opportunity that we as a field have had to examine the assumptions of MR in the analysis of DNAm on complex traits, due to the high yield of mQTL and the frequent occurrence of multiple independent mQTL per probe. While some (most likely in the epigenomics or functional genomics fields) might have assumed the result, others (most likely in the epidemiology / common disease fields, where much of the 450k DNAm analysis is done) assumed the converse, and we are simply trying to present evidence that speaks to the question.
>>> Ok

In regard to common and non-haematologically driven diseases cannot the response statement that “These DNAm sites are not generally on the causal pathway mediating genetic effects on disease” be stronger and state no causal evidence? What are the generalisable implications of the weakness of the causal results (line 561) where only just over half were in the correct direction? What do the authors consider this implies when this methodology is performed for single traits – and again when considering the potential number of mQTL in the genome.
• The method that we advance here is to look for a general relationship between the observed causal effects from mQTL across multiple probes, and the expected relationship under a model of causality. The power of the study is to enable this where previously it was not possible. But the limitation is that it only speaks to the general relationship and not to the relationship at any specific probe. The point, however, is that often it is the mQTL instrumentation design itself that is used to make a claim about probe-specific causal relationships, and so the design of this analysis helps to indicate the limitations of that approach without an appropriately large number of independent instruments for a particular probe.
>>> These limitations discussed should be included the manuscript

Reply: We now clarified this in the main text and in the Implications:

“We cannot determine with certainty the causal relationship of any specific site with a trait. To test if there was a general trend of DNAm sites causally influencing a trait we evaluated if the MR effect estimate based on the colocalising signals were consistent with those obtained based on the secondary signals."

“Together, across the sites that were analysable in this manner, these results indicate that those blood measured DNAm sites that have shared genetic factors with traits cannot be typically thought of as mediating the genetic association to the trait "

“Though we found substantial sharing of genetic signals between DNAm sites and complex traits, we were able to demonstrate that this was not predominantly due to DNAm variation being on the causal path from genotype to phenotype. While our results include <2% of the DNAm sites in the genome and are limited by the two-phase design, these findings have several implications especially in the context of EWAS studies that are often based on the same tissue and DNAm array.”

Clear effects of traits influencing DNA methylation in blood, such as Dekkers et al triglyceride DNAm associations, are supported mechanisms (line 619). As the causal model is not supported (line 570), it would be highly useful for the common disease field if the authors can use the power of this study to state whether the conclusion of previous EWAS papers cited here, like Wahl et al. – which identified predictive findings in pathogenically non-relevant tissues as well as consistent findings across multiple cells, are robust? Or are these findings only biomarkers of trait-related changes and genetic confounding?
• We do conclude, having investigated several large scale EWAS studies, that “These results indicate that traits causally influencing DNAm levels in blood is the most likely mechanism that gives rise to these EWAS hits.” (end of section on “Influence of traits on DNAm”). This result is echoed in the Implications. “Overall our data and results have resulted in the most comprehensive atlas of genetic effects to date. We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis.“ In the broader context, Lappalainen & Greally35 suggest that EWAS have been carried out to find loci that indicate cellular reprogramming. We found in this paper and others3,36, meaningful interpretation of EWAS findings is hampered by confounding factors and reverse causation. We agree that many EWAS studies do not account

for genetic confounding or test for reverse causation. If we as a field are to find loci that indicate cellular reprogramming, then we should in addition to genetic confounding also take into account other confounding variables such as transcriptome and generate new reference panels to adjust for cellular heterogeneity. The need for second generation EWAS is now apparent in the EWAS field and we agree that our mQTL catalogue would be part of this as we anticipate that some previously reported EWAS associations are likely due to reverse causation. As EWAS currently stands, at best it uses DNAm sites as a marker for relevant regulatory functions, without being able to speak to function.
To address the issue of genetic confounding in EWAS large meta-analyses are needed and we will investigate this in a separate manuscript. With regards to reverse caution, Wahl et al.16 did show that DNAm can be altered in response to increased BMI. Dekkers et al.17 also showed that DNAm was changed in response to altered lipid profiles.
>>> The authors state above "to address the issue of genetic confounding in EWAS large meta-analyses are needed". Is this not what this publication is? – and exploring these genetic effects is a core justification for this publication. Yes, both Dekkers et al. and Wahl et al. show the physiological impact of high lipids on the blood methylome, directly and indirectly, respectively. However, Wahl et al. controversially implied that a CpG's DNA methylation change in peripheral blood was 'causal' in obesity– not just the signature of the phenotype-related hyperlipidaemia, hyperglycaemia, inflammation, etc. as others have found. Furthermore, Wahl et al. identified consistent DNA methylation changes across multiple cell types, which is not coherent with our knowledge that biological changes with the haematological system impact differently on particular blood cell-types [You et al. Nat Comms 2020]. The authors should use this large meta-analysis to help advance the field by being more definitive about previous results where they can.

Reply: We agree with the reviewer that we should use large EWAS to advance the field. Wahl et al concluded that the 187 sites from EWAS were due to reverse causality (I.e. BMI influences those CpG sites) as shown in Figure 2 of Wahl et al. Our analysis supports this, we state that genetically instrumented BMI, when tested against those 187 sites, has larger associations than expected by chance (lines 704-706). Similar results were found for sites found in a triglycerides EWAS. In the text we conclude: "These results indicate that traits causally influencing DNAm levels in blood is the most likely mechanism that gives rise to these EWAS hits." In terms of whether this publication is a solution to confounding in EWAS – we do not make this claim. EWAS confounding could often occur due to large genetic influences, but there could also be small genetic influences not mapped in this study, or non-genetic confounding also. We have revised the two sentences of the Implications to reflect this:

"While our results include <2% of the DNAm sites in the genome and are limited by the two-phase design, these findings have several implications especially in the context of EWAS studies that are often based on the same tissue and DNAm array."

"We expect that this atlas will be of use to the scientific community for studies of genome regulation, contribute to the control of confounding in EWAS and to perform causality analysis."

With regards to the finding in Wahl et al where DNA methylation changes are shared across multiple cell types, this might be that the 187 sites were originally identified in bulk tissue comprising diverse cell types. You may therefore miss the cell type specific associations, especially associations for rare cell types and you are more likely to pick associations that are shared across cell types and tissues. This has also been shown in You et al. 2020 eg. hypomethylated smoking associated sites originally identified in whole blood (from Zeilinger et al.) showed a trend for hypomethylation in both myeloid and lymphoid lineages (Fig 1g in You et al.) but additional smoking associations in the myeloid lineage were found.

We now clarified this in the Implications by:

"However, as disease relevant signals and regulatory regions may be cell type specific, new analytical tools are required to infer cell type specific mQTL from bulk tissue."

The Selective analysis does not appear to be adding any substantial findings - as the conclusion given is only that a minority of regions containing DNAm sites have experienced positive selection and, significantly, no evidence is presented that the DNA methylation itself is driving any of the identified selection. This again also needs to be put in context that these data are only coming from 1.5% of the DNA methylome. Many additional SNPs not currently classified as mQTL would be if more CpG DNAm data was available.
• When performing the analysis to see if the mQTL were enriched for selection signatures, the question that we are trying to address is whether these variants, when compared against a background of variants with similar genomic properties, have higher overlap with genomic regions with selection signatures than expected by chance. Generally, an enrichment analysis does not have the complete set of causal variants available to its disposal (in a standard single trait GWAS, not all causal variants are known due to power), and the same is true here – we do not have all causal variants for DNAm. Our analysis reliably concludes that there are more selection signatures colocalising with our mQTL than we'd expect by chance, and that it is the fact that the positions associate with DNAm that drives the association (rather than other properties such as gene proximity, MAF, LD, etc). We feel that this is a reasonable and widely adopted approach to examining a set of discovery loci and shouldn't be disregarded just because don't have knowledge of every mQTL. The results are still relevant to the mQTL that we do have – and that is a vastly larger number than has been recorded previously.
>>> Yes, the point that was raised wasn't to disregard the findings but for the authors to extend what the implications of their findings could mean.

Reply: We thank the reviewer for clarifying this further. In response to the original comment we now changed the following:

"To summarize, our results provide the first evidence that selection may have shaped the landscape of DNAm values across the genome although the mechanism for the selection signals that exist at these loci remains unknown.**"**

to

"To summarize, our results provide the first evidence that selection may have shaped the landscape of DNAm values of the 450k sites although the mechanism for the selection signals that exist at these loci remains unknown.**"**

In the Implications we added:" While our results include <2% of the DNAm sites in the genome and are limited by the two-phase design, these findings have several implications especially in the context of EWAS studies that are often based on the same tissue and DNAm array."

As the authors state, the DNA sites are likely the primary target of selection - so the DNA methylation is just co-varying with genetic variation. In 19/42 at least one cis mQTL SNP with extreme SDS (Line 673) - but what would this be by chance with so many unidentified mQTLs? Can authors really say "these loci have been modified by disease pressure" (line 680) - with all

the missing data? Genetic selection has shaped the genome and due to the sheer number of mQTL will be represented in the observed DNA methylome even if the majority of DNAm is not functional change. How does this impact on their conclusions for height and cardiovascular disease? (line 683).
• We have tested the mQTL for enrichment of selection and so we know that our signal is not purely by chance. We agree with the reviewer that "the majority of DNAm is not functional change", though note that the case here in still unclear as we are considering the set of loci with the strongest selection signals. We agree that the interpretation of why these mQTL are enriched for selection signatures is hard, and we changed the text to clarify the difficulty in assigning mechanism to the selection signals that exist at these loci. Specifically, whilst it is the most likely explanation, we indeed don't know "these loci have been modified by disease pressure" and removed this claim.
We changed the text to: “A comparison showed that the genetic variance for cardiovascular disease associated mQTL or height associated mQTL with extreme SDS was higher when compared to all trait associated SNPs (Figure S49). To summarize, our results provide the first evidence that selection may have shaped the landscape of DNAm values across the genome although the mechanism for the selection signals that exist at these loci remains unknown.”
>>> Good to acknowledge

Regarding the Intermediate methylation values, the authors should acknowledge that the two isolated white-blood-cell subsets of T cells (CD4+) and Monocytes (CD14+) are not homogeneous entities and that further sub-types of CD4+ and CD14+ are clearly present and observable epigenetically. Therefore, cell-to-cell variability (line 358) cannot be definitively concluded and cell-type differences may still be present. These intermediate loci could, in fact, fit with loci that are enriched for important functional units that are specifically driving or represent these divergences into further subcell-types.
• We thank the reviewer for this comment and changed this sentence to: “By replicating this finding in two isolated white-blood-cell subsets(Figure S22), we showed that this is due to cell-to-cell variability16,37 or sub cell type differences which may indicate that these loci contribute to the divergence into further sub cell types.”
>>> Good to acknowledge

In regard to the Authors reply re results "support a mixed genetic architecture of small genome-wide polygenic effects and larger cis effects." The statement that “relatively little left to be discovered in cis” – it is unclear how the authors can state this if only assessing 1.5% of the DNA methylome and genetic variation is currently only limited to variants well-tagged by common SNPs (much structural variation, repetitive genome, and poorly sequenced regions of the genome remains). ~50% of the all CpGs in the human genome reside within the Repetitive genome that is very poorly covered by arrays for technical reasons.
• Again, we apologise that the wording in the text may have strayed from emphasising that the 450k array only represents a small proportion of the methylome. We have changed the text to clarify that we only assessed DNAm sites on the 450k array: “We then investigated how much of the heritability of variable DNAm can be explained by our mQTL associations on the 450k array using family-based heritability studies of DNAm33,38”
>>> Good to acknowledge

Furthermore, is not stating that the genetic architecture of DNAm is ‘polygenic’ rather obvious, as no previous experimental or genome-wide evidence implied it was monogenic or would fit with its epigenetic role? This meta-analysis points to a likelihood that mQLT will exist for almost all CpGs, except those few that are squarely within high dense CGI that are almost always unmethylated.

• Although this seems obvious no other study had the power to conclude this or focused on cis mQTL only or trans mQTL of GWA SNPs.
>>> Ok

Further to this, the authors state the “very small genetic effects on DNAm is valuable for gaining power to evaluate whether its variation has a substantial causal role in disease and other biological processes”. Conceptually the authors have not commented on the implications of their findings if 45% of array probes are potentially influenced genetically by mQLTs. Even extrapolating at this level implies potentially 12 million CpGs, but additionally, they estimate that with each probe there was an increase of 0.76 cis mQTL and 0.05 trans mQTL – which could push this up to a level of ~21 million CpGs being genetically influenced. This fits with the fact that this 450k array is more biased towards those CpGs nested within very high CpG dense CpG islands regions – which show extreme lack of variability (unmethylated on any allelic background - as confirmed in Figure 1b) so that the level of genetic influence of even ~45% is an underestimate. Therefore, it is clearly very easy
to find mQTL even in this meta-analysis of these reasonably homogenous population groups.
• We agree that many probes had low variability. The implication here is that a genetic effect will be harder to detect because the signal to noise ratio in the probe assay is lower, so the statistical power is lower. However, tranches of probes that have higher variability are not guaranteed to have more easily detectable genetic effects – they could be less genetic and more stochastic or unregulated. A small whole methylome sequencing study of twins5 demonstrated
that only a small proportion (~15–20% of 28M DNAm sites of the methylome) is differentially methylated in the population and that non-shared unique environmental factors are the main source of variation. Extrapolating the number of expected associations will depend on variance component analysis of probes that we do not have assayed (as DNAm sites will have variable genetic and environmental contributions) and is a worthy future study.
We moved the Implications paragraph about coverage to the Results:
“The microarray technology used in the majority of cohorts limited us to analyse only 2% of sites across the genome12, which are biased to promoters and strongly underrepresented regulatory elements. To explore the impact of expanding the coverage of arrays, we calculated the linear relationship between the median number of probes by gene on the 450k array and the median number of cis and trans mQTL. For each probe, we found an increase in the median of 0.76 cis mQTL ($p<9.03e-16$) and 0.05 trans mQTL ($p<1.47e-05$) (Figure S9). A similar increase was seen in non-genic regions. This indicates that expanding coverage will increase mQTL yield although this will depend on the genetic contribution of the DNAm site.”

>>> Ok – but would be good to state how many total CpGs could have mQTL even with the inadequate assumption of a similar rate. As discussed above, the authors should focus on what clear conclusions can be draw from their large array meta-analysis – not require support from less robust conclusions from an individual small low-coverage WGBS experiment.

Reply: We agree with the reviewer’s notion that it's impossible to extrapolate the number of mQTL beyond the 450k probes analysed. The specific analysis referred to by the reviewer here can only evaluate the density of mQTLs per CpG density according to this set of probes, and speculation beyond that is perhaps no more useful than the qualitative conclusion that there are more mQTL at untyped CpGs.

Additionally, future increases in power may bring more nuanced and further independent signals from their current LD clumped cis effects (without definitive knowledge that all effects can be completely coalesced down to only 1 variant). What are the implications regarding the inferences of causality if the vast majority of the DNA methylome has genetic influence? What does that mean in terms of the interpretation of MR with methylation?
• We added the following sentences in the Implications to discuss this:
"Fourth, DNAm might be on the causal pathway in a disease-relevant cell type or context."
..
"Finally, EPIC arrays2 and promising low-cost sequencing technologies19 offer an opportunity to make detailed interrogations of enhancer and other regulatory regions, which may be more fruitful in finding causal relationships with complex disease."
>>> Therefore, the authors need to include directly a comment on caution due to the potential for over-interpretation.

Reply: We added the following to the Implications:

"While our results include <2% of the DNAm sites in the genome and are limited by the two-phase design, these findings have several implications especially in the context of EWAS studies that are often based on the same tissue and DNAm array."

…

"Future studies may be more fruitful in finding causal relationships with complex disease. Either EPIC arrays[13] or low-cost sequencing technologies[69] will expedite detailed interrogations of enhancer and other regulatory regions. Especially, single molecule long-read sequencing promises to expand the genetic and epigenetic spectrum by allowing the detection of complex genetic variation such as allele specific DNAm and structural variation and different types of DNA modifications. However, as disease relevant signals and regulatory regions may be cell type specific, new analytical tools are required to infer cell type specific mQTL from bulk tissue. Given our projection of mQTL yields expected for future studies, pleiotropy involving mQTL is likely to be increasingly important to model when interpreting genotype-trait pathways."

How could this contribute to the development of second-generation EWAS as they state? (Line 730).
• We agree with the reviewer that this should be clarified. We changed this sentence to: "We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."
>>> Ok

Also, what is the potential of this increased power to detect more subtle technical variation - as well as direct genetic effects identified? Therefore, as the potential cis effect in the DNA methylome is extremely large, how can the authors definitely state that "Yet the majority of genetic effects are likely to act in trans?"
• We agree with the reviewer that these are important questions and they do naturally stem from the results that we have presented. We hope that our study can help shift the conversation in that direction, and that it is clear that the field as a whole needs to endeavour to explore these questions, and given how packed the paper is already it is difficult to extend the scope further."

>>> The authors need to more clearly state what they mean by this – re effect size of cis versus trans and how the current evidence implies the majority of genetic effects act in trans – also taking into consideration the relative genome space size of cis versus trans.

Reply: This very question is addressed directly in Gaunt et al 2016, which we cite as a motivator for the need for larger sample sizes to detect small trans-effects. In Gaunt et al., we used SNP heritability for each probe to estimate the proportion of the variation in a probe that is due to genetic variation. The SNP heritability was estimated for each probe using two variance components, the first generated using only SNPs within ±1 Mb of the CpG site (the *cis* component) and the second generated using all remaining SNPs (the *trans* component), such that var(y) = var(g_cis) + var(g_trans) + var(e), where the genetic variance due to all SNPs var(g) = var(gi_cis) + var(g_trans), and var(g)/var(p) is the total SNP heritability for a probe at a particular time point. This analysis showed that the proportion of *trans* heritability explained by *trans* mQTL is much smaller than the proportion of *cis* heritability explained by *cis* mQTL. We now clarified this in the paper.

"To date, only a small fraction of the total genetic variation estimated to influence DNAm across the genome has been identified, primarily restricted to cis mQTL[6]"

to

"To date, only a small fraction of the total genetic variation estimated to influence DNAm across the genome has been identified[7] and the proportion of *trans* heritability explained by *trans* mQTL (defined as more than 1Mb from the DNAm site) is much smaller than the proportion of *cis* heritability explained by *cis* mQTL."

The authors consider that this meta-analysis approach brings unique insights through increased power. Therefore, it would be good to convey more strongly how the synthesis of these data could drive the field forward, rather than supporting conclusions from older papers - many of which will now have been superseded due to recognised potential technical issues.
• Thank you for making this point. We feel that an important conclusion of this paper is not being assimilated – additional power to detect genetic effects has led to us attempting to understand how perturbation of the methylome leads to disease and functional variation, and we have found little in terms of these relationships. The major biological finding is that variation in the methylation sites we assayed does not have obvious functional effects for disease and either the genotype-phenotype map is far more complex than we currently understand, or DNAm at these sites in bulk tissue should be de-prioritised in understanding disease variation. Whether DNAm generally is on the causal molecular pathway to disease, when interrogated in a cell-type specific manner and having accounted for wide ranging pleiotropic possibilities, is not known. We have now updated the Implications and added:
"We expect that this atlas will be of use to the scientific community for studies of genome regulation and for EWAS to control for genetic confounding and to perform causality analysis."

>>> Yes, this is in fact the point in terms of driving the field forward - that as the author state, there are "not obvious functional effects for disease". Therefore, the request was to interrogate previous results and determine if their potential claims of functionality are incorrect.

Reply: Broader knowledge of genetic influences on DNAm has been postulated as a means to improve interpretability of EWAS, however as we stated above there are multiple avenues of confounding that cannot be completely solved by straightforward interrogation of mQTL catalogues. This has also been discussed in Chen et al. 2016 and later in Lappalainen and Greally 2017 where they envision that WGBS, ATACseq, RNAseq, scRNAseq assays are needed to allow the interpretation of EWAS. Method development and empirical studies are required to determine the most effective ways to utilise mQTL and other regulatory features information to factor out confounding, which is beyond the scope of this

paper. Also DNAm might be on the causal pathway in a disease-relevant cell type or context and better deconvolution methods to infer cell type specific mQTL from bulk tissue remains to be developed.

Also, it would be good to try to integrate the latest functional insights to the DNA methylome such as the site-specific tendencies of site-specific dependencies of DNMT and TET activity (Ginno et al Nature Comms 2020) and specific repressive and active roles with specific transcriptions factors (Yin et al Science 2017, Domcke et al Nature 2015) etc.
• We now introduced the TF section as follows:
"Recent studies have uncovered multiple types of transcription factor (TFs)/DNA interactions with DNAm including the binding of DNAm-sensitive TFs24-26. Epigenetic editing studies have revealed that local methylation and demethylation activities are affected by TF binding and cooperativity between TFs25,27."
>>> Good to acknowledge

Additional Major Points
1. "The role of inter-individual variation in DNA methylation on disease mechanisms is not yet well characterised" – Incorrect statement, the authors are referring to 'common' variation only and should clearly specify this – robust pathogenetic involvement of DNAm in rare disease such as Fragile X, Facioscapulohumeral muscular dystrophy, MLH1-deficient sporadic colorectal cancers, etc.
• We changed the text to: "The role of common inter-individual variation in DNA methylation (DNAm) on disease mechanisms is not yet well characterised."
>>> Good to clarify

2. The authors only cite papers re genetic influence on DNA methylation from 2014 onwards – but exclude landmark earlier work from Kerkel et al. Nature Genetics 2008 and others. Furthermore, the McRae et al. citation is problematic as they confused genetically-driven DNA methylation differences with the distinct biological mechanism of "transgenerational" epigenetic effects (only robustly seen in plants - Horsthemke N Comms 2018) – by including "transgenerational" in their title.
• Thanks for these comments. We have removed the McRae et al. citation from the following sentence "Because genetic influences on DNAm in blood have been shown to be widespread33,34". We do use BSGS heritability estimates later on from the McRae paper38 which we don't feel are controversial. Kerkel et al.43 (which is indeed a landmark study) describes allele specific methylation which is a comparison of alleles in a heterozygous individual whereas we refer to the association between genetic variants and DNA methylation in multiple individuals.
>>> It would be preferable to use other heritability estimates and not to cite and further propagate in the literature this erroneously titled paper in what will be a high-profile consortium paper. This fundamental error regarding 'transgenerational' has already caused significant confusion in the field. The Kerkel et al. paper was the first to show in humans the relationship between SNPs and methylation, by allelic variation, which is one of the drivers of the population variation observed.

Reply: We have now cited Shah et al. 2014 instead. This paper also describes the BSGS heritability estimates in detail. We also cited Kerkel et al. at the intro of the main text:

"Because genetic influences on DNAm in blood have been shown to be widespread[2-4], a powerful avenue into researching the functional consequences of changes in DNAm levels is to map genetic differences associated with population-level variation, identifying DNA methylation quantitative trait loci, (mQTL) that include both local (*cis* mQTL) and distal (*trans* mQTL) effects."

3. A conditional analysis within the locus of each mQTL identified 758,130 putative independent variants (line 295). Again, this needs to be put into context that only 1.5% of the DNA methylome is being measured, that median 2 and IQR 4 independent variants of just these CpGs and therefore how large the potential genetic to methylation space is.
• We moved the paragraph describing the limitation of the array from the Implications to the results to clarify this.
"The microarray technology used in the majority of cohorts limited us to analyse only 2% of sites across the genome12, which are biased to promoters and strongly underrepresented regulatory elements. To explore the impact of expanding the coverage of arrays, we calculated the linear relationship between the median number of probes by gene on the 450k array and the median number of cis and trans mQTL. For each probe, we found an increase in the median of 0.76 cis mQTL ($p<9.03e-16$) and 0.05 trans mQTL ($p<1.47e-05$) (Figure S9). A similar increase was seen in non-genic regions. This indicates that expanding coverage will increase mQTL yield although this will depend on the genetic contribution of the DNAm site."
>>> Good to acknowledge

4. The authors state that the cohorts have diverse geographical origins (line 689) – but these are only European ancestry population cohorts. What implication do their findings have regarding potential genetic effects in studies performed across population ancestry groups or with mixed population groups?
• We think it is vitally important to include non-European studies. Given the technical challenges of this study we wanted to restrict to a single ethnic group to avoid potential confounding that would be difficult to guard against. With this study as a reference we are now well positioned to analyse multi-ethnic studies and going forward this will be a major priority.
>>> Therefore 'diverse geographical origins' is not correct

Reply: The studies originate from multiple European countries, America and Australia. Therefore the use of diverse geographical origins is correct – We did not claim diverse ancestral origins. We completely agree that inclusion of diverse ancestries is crucially important, and work is underway for a forthcoming paper.

5. Are these results impacted on by the identification of further 450k array probe technical variability issues as described by Sugden et al. Patterns 2020?
• We used two approaches to exclude DNAm sites from our analyses. First, we excluded 50,186 DNAm sites that were masked by Zhou et al.44 which includes probes with potential cross reaction and probes that could not be mapped to genome. Secondly, we removed an additional 14,882 probes including multimapping probes (bisulfite converted sequences allowing two mismatches at any position mapped to the hg19 primary assembly) and probes with variants (MAF >5%, UK10K) at the CpG dinucleotide or the extension base (for type I probes).
Overall, we found that the reliability estimates from Sugden et al.45 were slightly lower for probes that we analysed (median = 0.196 vs median = 0.082). We then compared the reliability estimates to the five categories described in Zhou et al.44 and found that probes that were affected by a SNP have high reliability estimates (median= 0.502 vs median = 0.09). We also extracted reliability estimates for our mQTL probes and found that the reliability was much higher (median= 0.266 vs median = 0.0356) indicating that the type II error was small (which makes a lot of sense, since you can't have replication between cohorts if you can't measure the probe reliably). However, we may have missed genetic effects on unreliable probes.
>>> Ok, good to incorporate this new analysis

6. Statements such as “DNAm is usually associated with gene repression, with loss of DNAm reflecting enhancer or gene activation” (line 339) are an oversimplification and need to be clearly caveated in regard to the promoter focus of this array and the authors’ interpretations – i.e. ignores important role in gene body methylation transcription and integration of the epigenome with chromatin modifications (Baubec et al Nature 2015 etc.), identification of TF that require DNA methylation for binding (Yin et al Science 2017) etc.
• We have updated the wording with findings from these papers: “Recent studies have uncovered multiple types of transcription factor (TFs)/DNA interactions with DNAm including the binding of DNAm-sensitive TFs24-26. Epigenetic editing studies have revealed that local methylation and demethylation activities are affected by TF binding and cooperativity between TFs25,27.”
>>> Good to acknowledge

7. Authors need to state where findings are as expected and have potential clear explanations – e.g. ‘Sites that overlap TFBS are relatively hypomethylated’. TBFS are enriched within high CpG density loci (i.e. CpG islands) that are generally always unmethylated and ‘TFBS enrichments were not tissue specific’ also fits with these residing within CpG islands – which clearly overlap with promoters which are less tissue-specific.
• We have highlighted expected findings in the text:
“However, we found that the highest correlations were for associations involving trans-only sites (Adipose rb=0.92 (se =0.004); Brain rb=0.88 (se=0.009)) despite having on average smaller effect sizes than cis only associations, implying that they are less tissue specific than cis effects (Adipose rb=0.73 (se =0.002); Brain rb=0.59 (se=0.004)) which is line with the notion that promoters are less tissue-specific.”
…
“We found strong enrichments for the majority of TFs amongst DNAm sites with a trans association (cis+trans: 55%; trans only: 80%; cis only: 18%) which is in line with the observation that loss of DNAm at promoters is usually associated with gene activation30, and amongst cis-acting SNPs (cis only: 96%, cis+trans: 91%, trans only: 1%) (Figures 2B, S28, S29). Consistent with the observation that trans only DNAm sites are enriched for CpG islands (Figure S19), sites that overlap TFBS were relatively hypomethylated independent of tissue (weighted mean DNAm levels = 21% vs 52%, $p<2.2e-16$) (Figure S30) and we found that generally the TFBS enrichments were not tissue specific (Table S10-11, Figure S28-29).”
>>> Good to acknowledge this known biology

8. The authors cite an old paper for support that “trans-SNPs were typically found in intergenic regions which are depleted for complex trait loci” from Schork et al. PLoS Genetic 2013. Are this paper’s conclusions consistent with current knowledge of predominately noncoding GWAS findings - as was summarising early arrays used for GWAS from 2010/2011?
• We agree with the reviewer. Approximately 90% of the phenotype associated SNPs identified by GWAS lie within non-coding regions. After accounting for LD, enrichment analyses have shown that GWA SNPs are enriched for coding, conserved and regulatory regions such as promoter and enhancer regions 46(Finucane et al. 2015). We didn’t find any enrichments for these regulatory annotations for trans only SNPs.
“These enrichments correspond to the results reported earlier, in which trans-SNPs were typically depleted for enhancer and promoter regions, whereas complex b btrait loci are enriched for coding and regulatory regions47.”
>>> Ok, good to clarify

9. “1,373 putative examples in which at least one DNAm site putatively shared a genetic factor

with 71 traits including only 19 disease (line 540). Again, how do authors interpret this if genetic influences on DNAm are ubiquitous?
• This doesn't seem to be a contradiction – not every variant will influence DNA methylation, and not every variant will influence complex traits, and the overlap doesn't need to be complete?
>>> The comment was not regarding this being a contradiction and the requirement for complete overlap – but with so many SNPs already impacting on observed CpGs and the implications of all the unassayed CpGs - and what this means for the potential over-interpretation of current positive findings.

Reply: We're not sure we can state any solid implications about the ubiquity of mQTL and colocalization. We can generally speak to the point that observed pleiotropy will likely increase. We have added the following statement to the Implications:

"Given our projection of mQTL yields expected for future studies, pleiotropy involving mQTL is likely to be increasingly important to model when interpreting genotype-trait pathways."

10. Regarding the GWAS signal and DNAm site colocalised analysis (line 545). What is the true relevance of this closest determination – as the authors don't have exhaustive CpGs data in the region so may have no data other close CpGs and, furthermore, the GWAS signal unlikely to be the actual causal functional SNP in each locus?
• Coloc doesn't require knowledge of the causal variant, it just looks for consistent pattern of association at a locus across two traits. We likely don't detect every instance of a shared causal variant, but we only make inference about the ones that we are able to.
>>> However, the authors need to acknowledge the large number of missing CpG data in this array analysis.

Reply: Unmeasured CpGs that are closer to the mQTL could be more or less likely to colocalize, we don't know until those data become available. But the point of the analysis is that for the 450k array data we shouldn't assume the closest CpG to a GWAS trait is the most important. We have edited as follows:

"Notably, in only 18.1% of the cases where a GWAS signal and an assayed 450k DNAm site colocalised, was that DNAm site the closest DNAm site to the signal."

11. For the widespread changes in DNAm levels (line 599) can the authors give an estimate of how many CpGs are potentially contributing for each these 5 outlier traits?
• The manner in which this was calculated does not allow us to count the precise number of CpGs that are influenced by these traits, as we are comparing the distribution of observed test statistics against the theoretical distribution under the null hypothesis of no associations.
>>> Ok

Reviewer #3:
Remarks to the Author:
The authors have addressed all my comments. I have no additional comments.

Our flexible approach during the COVID-19 pandemic
If you need more time at any stage of the peer-review process, please do let us know. While our systems will continue to remind you of the original timelines, we aim to be as flexible as possible during the current pandemic.

This email has been sent through the Springer Nature Tracking System NY-610A-NPG&MTS

Confidentiality Statement:

Decision Letter, first revision:

Our ref: NG-A56091R

15th Dec 2020

Dear Dr. Min,

Thank you for submitting your revised manuscript "Genomic and phenomic insights from an atlas of genetic effects on DNA methylation" (NG-A56091R). I am delighted to say that we will be happy in principle to publish this in Nature Genetics, pending minor revisions to comply with our editorial and formatting guidelines.

** Note that we will send you a checklist detailing these editorial and formatting requirements in about a week. Please do not finalize your revisions or upload the final materials until you receive this additional information.**

In recognition of the time and expertise our reviewers provide to Nature Genetics's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "Genomic and phenomic insights from an atlas of genetic effects on DNA methylation". For those reviewers who give their assent, we will be publishing their names alongside the published article.

While we prepare these instructions, we encourage the Corresponding Author to begin to review and collect the following:

-- Confirmation from all authors that the manuscript correctly states their names, institutional affiliations, funding IDs, consortium membership and roles, author or collaborator status, and author contributions.

-- Declarations of any financial and non-financial competing interests from any author. For the sake of transparency and to help readers form their own judgment of potential bias, the Nature Research Journals require authors to declare any financial and non-financial competing interests in relation to the work described in the submitted manuscript. This declaration must be complete, including author initials, in the final manuscript text.

If you have any questions as you begin to prepare your submission please feel free to contact our Editorial offices at genetics@us.nature.com. We are happy to assist you.

Thank you again for your interest in Nature Genetics.

Congratulations on the paper!

All the best,

Catherine

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071

Final Decision Letter:

In reply please quote: NG-A56091R1 Min

12th Jul 2021

Dear Dr. Min,

I am delighted to say that your manuscript "Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation" has been accepted for publication in an upcoming issue of Nature Genetics.

Prior to setting your manuscript, we may make minor changes to enhance the lucidity of the text and with reference to our house style. We therefore ask that you examine the proofs most carefully to ensure that we have not inadvertently altered the sense of your text in any way.

Once your manuscript is typeset and you have completed the appropriate grant of rights, 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.

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-A56091R1) 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

<B>Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, 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 our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard 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 Research offers an immediate open access option only for papers that were first submitted after 1 January, 2021.

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

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 <a href="https://www.nature.com/reprints/author-

reprints.html">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-A56091R1). Further information can be found at https://www.nature.com/nprot/.

Congratulations on the paper!

All the best,

Catherine

Catherine Potenski, PhD
Chief Editor
Nature Genetics
1 NY Plaza, 47th Fl.
New York, NY 10004
catherine.potenski@us.nature.com
https://orcid.org/0000-0002-4843-7071