nature portfolio

Peer Review File

A compendium of genetic variations associated with promoter usage across 49 human tissues

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this study, Yuan and Tong et al. introduce a framework for identifying genetic variants that influence alternative promoter activity and selection, named promoter activity quantitative trait loci (paQTLs) and promoter usage quantitative trait loci (puQTLs). The findings demonstrate that puQTLs play a crucial role in uncovering new targets and regulatory patterns. In my opinion, this study is comprehensive and well-written. Understanding the genetics of promoter selection should contribute to the functional interpretation of genetic variants and their roles in disease risk. However, there are a few areas that require further clarification and improvement.

Major comments:

1. The authors should explain the significance of focusing on the relative usage of promoters instead of absolute activity in biological contexts. In this case, their interest lies in understanding how genetic variants influence the relative abundance of multiple isoforms of a single gene. It is important to clarify why understanding the relative usage of promoters is crucial for comprehending complex biology. Moreover, it would be helpful to discuss what could happen if the balance of promoter activity is disrupted. Providing examples of scenarios where the relative usage of promoters is worth discussing, as opposed to absolute activity, would further enhance the explanation.

2. I think it is important to check the consistency between this paper's results and previous reports' one. The authors should compare the paQTL or puQTL in EBV-transformed lymphocytes with previously reported CAGE-based puQTL (Garieri et al., 2017) and RNA-seq-based puQTL (Kubota and Suyama, 2022) since they mapped puQTL in the same cell types.

3. How many puQTL are statistically fine-mapped with significant confidence? The fine-mapping would be useful to discuss causality between genetic variants and promoter activities. The authors are trying to associate a specific genetic variant and the target promoter by looking at epigenetic marks and chromatin interactions (i.e. rs487174 - Promoter 21007), in which case fine-mapping would help them validate the functional connections.

4. It is unclear how the accuracy of the estimated promoter activities is confirmed. The authors should provide a more detailed explanation of the assessment method for accuracy to enhance readers' understanding.

5. Which cell types were used for the HiChIP data? It is crucial to select suitable cell types as local chromatin interactions are often specific to particular cell types. The authors should provide clarification on which HiChIP data were used for each tissue in Figure 1F in the Methods section.

6. How many puQTLs are not mapped as paQTLs of prominent promoters in each tissue? Please provide the number or percentage to emphasize the importance of focusing on relative usage of promoters rather than absolute activity.

7. In Figure 5D and E, the authors demonstrate that rs487174 is linked to Promoter 21007 as a puQTL but not as a paQTL. Given this, is this genetic variant associated with Promoter 21008 as a paQTL?

8. In Figure 5K, the authors demonstrate the similarity of the effect size in puQTL and sQTL (rs487174 - PRDX1). However, it remains unclear which introns or exons are associated with the genetic variant as sQTL. Specifically, it is important to clarify whether it is the first intron of PRDX1 because the

calculation method of promoter activity employed in this paper seems to be based on junction read counts on the first intron of the transcript, in which case inclusion ratio of the first intron would be almost the same thing as promoter activity. In this case, the comparison of puQTL and sQTL here would be just looking at the same thing. It would be beneficial for the authors to include the genomic coordinates and the number of the intron.

9. Regarding comment #8, it is important for the authors to provide information about the source of the sQTL data and explain how the mapping of those data was conducted. The method of sQTL mapping should be clearly described.

10. In Figure 3J and K, the authors demonstrate a significant depletion of eQTLs in promoters. Is this consistent with previous eQTL studies? Please cite some eQTL papers and compare the results.

11. In Figure S8D, the authors demonstrate that rs7738430 is linked to the activity of Promoter 68943 with a high beta value. However, the relationship does not appear to be linear, possibly because of the very low activity level in individuals with the aa genotype. Please verify the authenticity of this association and modify the figure to enhance readers' intuitive understanding.

Minor comments:

1. Related to major comment #4, please add a few sentences about how the promoter activity was estimated in this study.

2. Lines 81-87: I don't understand why these last two out of the three mentioned points are related to concerns in the definition of promoter usage. The structure of the sentence should be reorganized.

3. Fig. 1E: Please add an asterisk next to "Fraction of ~~".

4. Please cite the original paper of Bedtools at line 597 on page 26.

5. Fig. 5K: Please increase the size of the circles in the boxes.

6. Fig. 5C-D: How about prmtr.21008

7. Fig. 3K: It appears that the P-value is incorrect. Please correct it.

Overall, this manuscript has great potential, but further revisions are needed to address the aforementioned points. With these improvements, the study will make a significant contribution to the scientific community.

Reviewer #2 (Remarks to the Author):

This paper describes a new analysis of the GTEX dataset to identify promoter usage QTLs (puQTLs), a class of gene regulation that is distinct from previous reports investigating eQTLs or promotor activity in the GTEX dataset. The introduction is clearly written, and the authors have clearly justified why this additional layer of regulation is worth investigating. The authors identify puQTLs and perform multiple analyses to characterize the properties of puQTLs, how they compare to eQTLs and paQTLs, and identity potential ways in which puQTLs expand upon previous QTL discoveries in this well studied dataset. The paper and analysis represent a significant body of work. However, the methods in several

places are unclear, and several analytical points need to be clarified for the full significance of the findings to be interpretable.

Major Points

The authors describe several filtering steps prior to analyses, including removing promoters with missing value exceeding 50% per tissue and 80% per individual. As the final value used is a proportion, ie the usage of the major promotor compared to other promotors, it is unclear why these filters were needed, and particularly if removing them could inflate the final proportion of the major promoter in individuals where minor promotors were filtered. Could the authors please explain their justification for selecting these missingness thresholds, and detail what impact they have on the final results?

The reporting of the number of puQTLs is unclear. This is reported as 303,025 puQTLs across 48 human tissues. How many unique loci does this represent?

The text states that only the most commonly used promoter was selected for further analysis and further that "Specifically, puQTL focuses on genes, assigning a single promoter usage value to each gene and correlating it with genotypes... Unlike paQTL, which considers all promoters, puQTL identifies the most frequently utilized promoter to elucidate the correlation between genetic variations and promoter usage." However, in the text several mentions are made of genes with multiple puQTLs. It is unclear from the text how multiple puQTLs per gene have been identified, or what they represent. I would assume multiple puQTLs per gene were a result of conditional analyses to identify additional independent SNPs regulating the usage of the most frequent promoter? This is not detailed in the methods and text. The authors should explicitly explain how they defined multiple puQTLs per gene and add this to the methods.

The authors report the overlap betewen puQTL -gene pairs and paQTL pairs, and later with eQTLs. How were these overlaps calculated? Was only the lead SNP used for each dataset? Or were linked SNPs included as well? If so was there an LD threshold applied between the lead and linked SNPs in order to identify true overlaps? A description of the methods used to calculate the overlaps should be added to the methods.

In Figure 3F and 3G, the authors have split the dataset into directionally concordant subsets and then performed correlation analysis. This massively inflates the correlation and is not meaningful. Could the authors please display and report the overall correlation between eQTL effect size and puQTL size, including all examined pairs?

The authors report an overlap between the percent of genes with reported eQTLs and reported puQTLs. As the number of genes with an eQTL in the GTEX dataset (and eQTL datasets in general) is heavily dependent on sample size, surely the overlap here is mostly driven by the different sample sizes across tissues and subsequent power to detect QTLs rather than on any underlying molecular explanation?

The authors state "Remarkably, puQTLs displayed a significantly stronger correlation with GWAS traits in comparison to paQTLs and eQTLs. " But they do not clarify what evidence this statement is based on. Could the authors please clarify what evidence supports this strong statement? There are several potential relevant figures presented in the supplemental figure, but these are not clear in the text and I was not convinced that this statement is supported by evidence.

In the colocalization methods, the authors state: "To conduct the colocalization analysis of puQTL-

promoter pairs, we first obtained the significant SNPs from each GWAS trait (P-value < 5 x 10-8) and subsequently compared them with the nominal puQTLs (FDR < 0.05). " I do not think this approach is valid. The method they used, Coloc requires full summary statistics from both traits. Inputting only the GWAS significant hits and nominal puQTLs into COLOC will massively inflate the number of co-localizations. This analysis should be rerun with full summary statistics of both traits.

The authors do not include what methods/parameters they used for colocalization of the paQTLs or the eQTLs. This should be added.

The Data Availability statement does not detail how the results of this study – the puQTL summary statistics will be made available.

Reviewer #3 (Remarks to the Author):

In this paper, Yuan et al present an analysis of promoter usage QTLs (puQTL) in a subset of GTEx samples. While the paper presents some interesting results about the properties of puQTLs, its utility as a resource is strongly limited by using only a subset of GTEx samples, reliance on outdated gene annotations (GENCODE v19, released in 2013), and a complete lack of publicly available summary statistics. Furthermore, the introduction of the paper overclaims the novelty of the work by misrepresenting some of the previous work.

Major concerns
1. On lines 73-78, the authors make two claims: 1. "Prior investigations commonly evaluate promoter activity in a restricted set of tissues. However, there is a notable absence of a comprehensive study that examines promoter usage across a wide range of tissues.
2. "Besides, the strategy used in these studies is to measure promoter activity for each isoform by using the absolute expression level of promoter regions (19,20)."

Both of these claims are demonstrably false. First, these claims ignore the work by Alasoo, et al 2019 (1), which quantified the usage (not the activity) of each promoter relative to all other promoters of the same gene, thus exactly following the puQTL naming convention in this paper. While there are methodological differences in how promoter usage is quantified in this paper versus how it was quantified by Alasoo et al, it is not obvious that one of the approaches is superior to the other and no such comparison is presented in the paper. Secondly, the txrevise method developed by Alasoo et al to quantify promoter usage (along with internal exon usage and 3'end usage) has since been applied to all RNA-seq datasets present in the eQTL Catalogue (https://www.ebi.ac.uk/eqtl/) (2). These summary statistics have been publicly available since 2020 and have been continuously updated, most recently in April 2023 (https://www.ebi.ac.uk/eqtl/Release_notes/) (3).

Thus it is false to claim that previous studies have not quantified promoter usage (as opposed to promoter activity) and that this analysis has not been extended to a large number of tissues and cell types. The authors should rewrite the introduction of their paper to accurately reflect the published literature.

2. It is unclear how many RNA-seq samples were included in the analysis. The abstract mentions 503 individuals, which is significantly less than 838 donors present in the GTEx version 8 release (published in 202) (4). Furthermore, the GTEx v8 release contained data from 15,201 samples, but the number of samples analysed in this study is unclear as it is not mentioned anywhere in the paper. The method section only refers back to the paper by Demircioglu et al (5), which only analysed 6,674 GTEx samples (44% of GTEx v8). Is 6,674 the sample size used in this study? If yes, then the authors

should clearly state this early in the paper. They should also be explicit that they have used less than half of the data available from GTEx. Moreover, the latest release of the eQTL Catalogue contains uniformly processed puQTLs for 25,724 samples from 109 datasets, which is significantly larger than the current study.

3. The methods section currently does not provide any information on how the promoter activity values were computed from the GTEx RNA-seq data. Please add information about which software was used and with which parameters. Also, it would be important to justify why you decided to use transcript annotations from GENCODE v19 (released over 10 years ago) when much newer annotations are available. Finally, GENCODE v19 was based on GRCh37 coordinates while the WGS genotype data that you used was based on the GRCh38 coordinates. How did you reconcile this difference? Did you use liftOver or some other tool to convert the coordinates from one version of the reference genome to the other? If yes, then this should also be explicitly mentioned in the methods.

4. The title of the paper starts with "An atlas …" which implies that it describes a resource that could be used by other researchers. Is this true? If yes, then it would be important to make the summary statistics from the analysis publicly available. Currently, the data availability statement does not mention summary statistics at all.

References

1. K. Alasoo, J. Rodrigues, J. Danesh, D. F. Freitag, D. S. Paul, D. J. Gaffney, Genetic effects on promoter usage are highly context-specific and contribute to complex traits. Elife. 8 (2019), doi:10.7554/eLife.41673.
2. N. Kerimov, J. D. Hayhurst, K. Peikova, J. R. Manning, P. Walter, L. Kolberg, M. Samoviča, M. P. Sakthivel, I. Kuzmin, S. J. Trevanion, T. Burdett, S. Jupp, H. Parkinson, I. Papatheodorou, A. D. Yates, D. R. Zerbino, K. Alasoo, A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. 53, 1290–1299 (2021).
3. N. Kerimov, R. Tambets, J. D. Hayhurst, I. Rahu, P. Kolberg, U. Raudvere, I. Kuzmin, A. Chowdhary, A. Vija, H. J. Teras, M. Kanai, J. Ulirsch, M. Ryten, J. Hardy, S. Guelfi, D. Trabzuni, S. Kim-Hellmuth, W. Rayner, H. Finucane, H. Peterson, A. Mosaku, H. Parkinson, K. Alasoo, eQTL Catalogue 2023: New datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs. PLoS Genet. 19, e1010932 (2023).
4. The GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 369, 1318–1330 (2020).
5. D. Demircioğlu, E. Cukuroglu, M. Kindermans, T. Nandi, C. Calabrese, N. A. Fonseca, A. Kahles, K.-V. Lehmann, O. Stegle, A. Brazma, A. N. Brooks, G. Rätsch, P. Tan, J. Göke, A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. Cell. 178, 1465–1477.e17 (2019).

Point-by-point address of the comments raised by the reviewers for manuscript NCOMMS-23-45003-T

We would like to extend our gratitude to the editor and the reviewers for the thorough evaluation of our manuscript and for providing valuable feedback. We are delighted to report that all of the reviewers' comments have been meticulously incorporated in the revised version. The reviewers' insightful remarks and constructive suggestions have notably improved the overall quality of our work. Extensive revisions have been made to address the concerns raised by the reviewers. A detailed response to the reviewers' comments can be found below.

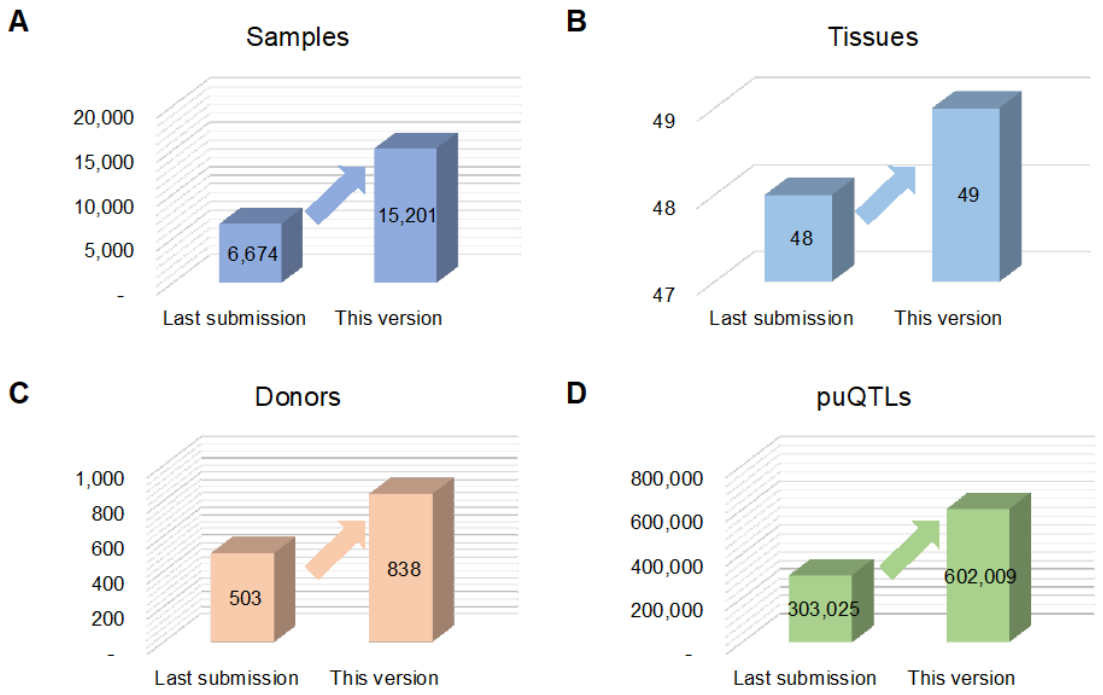
Reviewer #1 (Remarks to the Author):

In this study, Yuan and Tong et al. introduce a framework for identifying genetic variants that influence alternative promoter activity and selection, named promoter activity quantitative trait loci (paQTLs) and promoter usage quantitative trait loci (puQTLs). The findings demonstrate that puQTLs play a crucial role in uncovering new targets and regulatory patterns. In my opinion, this study is comprehensive and well-written. Understanding the genetics of promoter selection should contribute to the functional interpretation of genetic variants and their roles in disease risk. However, there are a few areas that require further clarification and improvement.

RESPONSE: We express our gratitude to the reviewer for recognizing the significance of our study and providing valuable feedback. We have thoroughly considered all the comments and incorporated substantial revisions into the manuscript, with major changes clearly indicated in red for ease of reference.

In the present revision, the introduction has been expanded to provide a more comprehensive rationale for the importance of elucidating promoter usage in the context of tissue development. Furthermore, **Response Figures 2-3** have been generated to specifically address the feedback provided by reviewer 1. The incorporation of supplementary figure panels, specifically **Figures S3C, S3D, S7, S8, S9, and S14**, is instrumental in presenting noteworthy discoveries that have arisen following the integration of feedback from reviewer 1.

Additionally, in response to the feedback from Reviewer #3, we have made substantial updates to the manuscript. Specifically, the GTEx RNA-seq dataset has been updated from version 6 to version 8, which now encompasses 15,201 samples from 838 donors across 49 tissues. Additionally, we have transitioned to the GRCh38 human genome reference and GENCODE v26 genome annotation file, aligning with the version utilized in GTEx v8. Notably, the sample size has expanded from 6,674 to 15,201 since the previous submission, thereby enhancing the power of our dataset for mapping puQTLs across 49 tissues (**Response Figure 1**). It is important to note that due to the increased sample size and updated genome annotation, certain details such as promoter ID, effect size values, and false discovery rate (FDR) may have been updated, leading to corresponding changes in the figures. Nevertheless, the overarching genome-wide conclusions remain unchanged and may have been bolstered by the heightened statistical power of our model. (**Page 3, Lines 35-37; Page 7, Lines 131-135**)



Response Figure 1. Comparisons of the number of samples **(A)**, tissues **(B)**, donors **(C)** and puQTLs **(D)** between last submission and this revised version.

Major comments:

1. The authors should explain the significance of focusing on the relative usage of promoters instead of absolute activity in biological contexts. In this case, their interest lies in understanding how genetic variants influence the relative abundance of multiple isoforms of a single gene. It is important to clarify why understanding the relative usage of promoters is crucial for comprehending complex biology. Moreover, it would be helpful to discuss what could happen if the balance of promoter activity is disrupted. Providing examples of scenarios where the relative usage of promoters is worth discussing, as opposed to absolute activity, would further enhance the explanation.

RESPONSE: We would like to express our gratitude for the valuable comments, which will contribute to enhancing the biological significance of relative promoter usage in contrast to absolute promoter activity. Our examination of previous studies has revealed that maintaining a balance of promoter activity is crucial for both tissue development and the dysregulation observed in cancer (Garieri M, *et al., Nat Commun.,* 2017; Demircioglu D, et al., *Cell*, 2019). For example, in the case of the gene *GJB1*, the selection of the 3'-most promoter is specific to CNS tissue, while the 5'-most promoter is selected in other tissues. Similarly, for the gene *CDK4*, the selection of the 5'-most promoter corresponds to a normal cellular state, whereas a disruption in the balance of promoter activity leads to the transformation of cells into cancerous states. The examples such as *GJB1* and *CDK4* reveals that while multiple alternative promoters may exist for a gene, typically only one are predominantly utilized in a certain tissue or cell type, and their relative usage follows an inverse pattern.

Current literature indicates that each gene typically possesses a dominant or optimal promoter, with alternative promoters playing a limited and potentially detrimental role (Garieri M, *et al., Nat Commun.,* 2017; Demircioglu D, *et al., Cell*, 2019). Consistently, in the instance of the gene *STAU2*, which has six alternative promoters, only one is highly expressed in basal cells, another is highly expressed in other subtypes, while the remaining four are not differentially expressed. Similarly, for the gene *TTC23*, which has three promoters, the promoter 1 and 2 exhibited opposite associations with the presence of genetic variant rs8028374, while the alternative promoter 4 is expressed at a low level and its proportion does not change with different genotypes. This observation leads us to consider that when only two

promoters are primarily selected, it may not be necessary to account for the activity of all promoters. Furthermore, if the proportion of the primarily selected promoters follows an inverse relationship, introducing the concept of relative promoter usage, where the sum of proportions equals 1, may effectively reduce the dimensionality of promoter selection compared to aggregating the activity of all promoters. We have incorporated the above discussion and examples into the Introduction and Discussion sections. (**Page 4, Lines 58-61; Page 5, Lines 85-94**)

2. I think it is important to check the consistency between this paper's results and previous reports' one. The authors should compare the paQTL or puQTL in EBV-transformed lymphocytes with previously reported CAGE-based puQTL (Garieri et al., 2017) and RNA-seq-based puQTL (Kubota and Suyama, 2022) since they mapped puQTL in the same cell types.

RESPONSE: We express our gratitude to the reviewer for the insightful feedback. In this study, we have conducted a comparative analysis of puQTLs and paQTLs in EBV transformed lymphocytes, in relation to previously reported puQTLs, which were identified through CAGE-based puQTL and RNA-seq-based puQTL methods. Given that both the CAGE-based and RNA-seq-based puQTLs in these published datasets are at the transcript-level, it is more appropriate to juxtapose the paQTLs identified in our study with those in the published dataset. Upon conducting our analysis, we identified 143,170 paQTLs in EBV transformed lymphocytes, a substantial increase compared to the 3,004 CAGE-based puQTL-gene pairs and 2,325 RNA-seq-based puQTL-gene pairs previously published. Notably, our study revealed that 9.95% of CAGE-based puQTL-gene pairs and 13.1% of RNA-seq-based puQTL-gene pairs were rediscovered, despite the disparate data sources (**Figure S3C**). Furthermore, we observed that less than 2% of puQTL-gene pairs overlapped between the two published datasets. This signifies the robustness and comprehensiveness of our findings in relation to previously established research. We have incorporated the above results into Results section. (**Pages 9-10, Lines 193-200**)

3. How many puQTL are statistically fine-mapped with significant confidence? The fine-

mapping would be useful to discuss causality between genetic variants and promoter activities. The authors are trying to associate a specific genetic variant and the target promoter by looking at epigenetic marks and chromatin interactions (i.e. rs487174 - Promoter 21007), in which case fine-mapping would help them validate the functional connections.

RESPONSE: We express our gratitude to the reviewer for providing valuable insights that will enhance our understanding of the causal relationship between puQTLs and promoter usage. Our study includes a meticulous fine-mapping analysis to ascertain the prevalence of puQTLs with causal effects across 49 different tissues. Notably, we observed that the median proportion of puQTLs with causal effects exceeds 75% in all examined tissues (**Figure S14A**), with an overall median proportion of 86% across the 49 tissues (**Figure S14B**). These findings strongly indicate the functional significance of puQTLs. We have incorporated these findings into the Results section of the revised manuscript. (**Pages 21-22, Lines 487-493; Pages 33-34, Lines 786-793**)

4. It is unclear how the accuracy of the estimated promoter activities is confirmed. The authors should provide a more detailed explanation of the assessment method for accuracy to enhance readers' understanding.

RESPONSE: We thank the reviewer for pointing out this issue. In the revised manuscript, we have provided a more detailed explanation of the promoter activity assessment method and made a clarification of accuracy of the method. We used a tool called *proActiv* to evaluate both the absolute and relative promoter activities using a method that counts junction reads. The junction read counts files were obtained from the GTEx portal as per the official instructions (https://gen3.theanvil.io/explorer), and we collected junction read files of 15,201 samples in the SJ.out.tab format. The version of the annotation files used in the *proActiv* is GENCODE v26, which matches the version used in GTEx v8 project.

The R package *proActiv* developed by Demircioglu et al., is designed to estimate promoter activity from RNA-seq data (Demircioglu D, *et al., Cell*, 2019). This approach involves quantifying promoter activity using a set of unique junction reads. To evaluate the performance

of *proActiv*, the developers conducted a comparative analysis of epigenetic modification signals near primary promoters. This analysis integrated previously published H3K4me3 histone data and CAGE data. The results indicated that promoter activity predicted by *proActiv* exhibits greater consistency with CAGE data and H3K4me3 histone data compared to other methods. Additionally, in a recent study, they also correlated predicted promoter activities from RNA-seq data with the captured transcriptomes identified from Iso-seq data, demonstrating that promoter activity inferred using *proActiv* is most consistent with the Iso-seq data (Huang KK, *et al., Genome Biol.,* 2021). These results validate the reliability of promoter activity estimation by *proActiv*. (**Page 7, Lines 128-131; Page 25, Lines 572-585**)

5. Which cell types were used for the HiChIP data? It is crucial to select suitable cell types as local chromatin interactions are often specific to particular cell types. The authors should provide clarification on which HiChIP data were used for each tissue in Figure 1F in the Methods section.

RESPONSE: We thank the reviewer for the constructive suggestions. We have taken into account the tissue specificity of chromatin interactions by incorporating data from HiChIPdb. In order to provide a comprehensive overview, we have included a table (**Table S3**) detailing the tissue of puQTLs and cell type of HiChIP data, along with the corresponding GEO accession number. These details have been incorporated into the Methods section for clarity and transparency. (**Page 7, Lines 143-144; Page 27, Lines 632-636**)

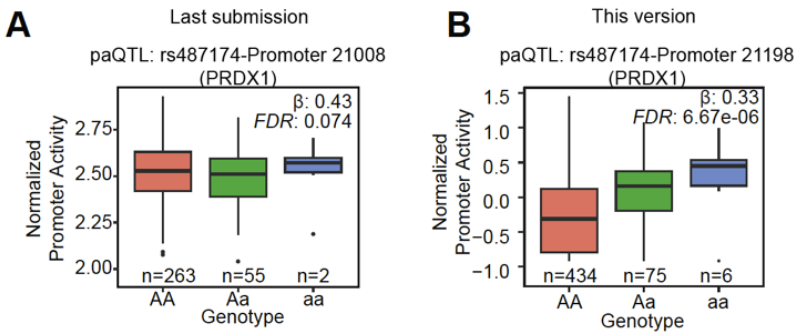
6. How many puQTLs are not mapped as paQTLs of prominent promoters in each tissue? Please provide the number or percentage to emphasize the importance of focusing on relative usage of promoters rather than absolute activity.

RESPONSE: Thank you for the insightful comments. In the revised manuscript, we have classified the significant puQTL-gene pairs into three types: the first type is that the puQTL-gene pair is not identified in the paQTL-gene pair set; the second type is that the puQTL-gene pair is also identified in the paQTL-gene pair set, but it is not statistically significant; and the

third type is that the puQTL-gene pair is also identified in the paQTL-gene pair set and with statistical significance. When we compared the proportion of three categories of puQTL-gene pairs, we discovered that around 75% of puQTL-gene pairs are not defined as significant paQTL-gene pairs across 49 tissues, highlighting the importance of focusing on relative usage of promoters (**Figure S3D**). (**Page 10, Lines 205-207**)

7. In Figure 5D and E, the authors demonstrate that rs487174 is linked to Promoter 21007 as a puQTL but not as a paQTL. Given this, is this genetic variant associated with Promoter 21008 as a paQTL?

RESPONSE: We express our gratitude to the reviewer for the insightful feedback. In our initial submission, we did not observe a significant association between rs487174 and promoter 21008 as a paQTL, as the false discovery rate (FDR) did not meet the predetermined statistical significance threshold of FDR < 0.1, with a calculated FDR of 0.074 (**Response Figure 2A**). Consequently, we did not present the association between rs487174 and promoter 21008 in our findings. However, in this revised version, we have redefined the IDs for promoter 21007 and 21008 to 21197 and 21198, respectively, in accordance with the updated GENCODE annotation used in our revised manuscript. Furthermore, the sample size of lung tissue has been increased from 320 to 515, thereby augmenting the statistical power of the paQTL mapping model. As a result, the FDR for the association between paQTL rs487174 and promoter 21198 now meets the FDR criteria, with a calculated FDR of $6.67e^{-06}$ (**Response Figure 2B**), and is therefore depicted in **Figure 5F**. Our revised results demonstrate a significant association between rs487174 and promoter 21197 as a puQTL, but not as a paQTL. Conversely, rs487174 is found to be significantly associated with promoter 21198 as a paQTL. (**Page 17, Lines 380-384**)



Response Figure 2. The relationship between the presence of paQTL rs487174 and promoter 21008 is explored in both the initial submission (A) and the revised version (B) of the manuscript.

8. In Figure 5K, the authors demonstrate the similarity of the effect size in puQTL and sQTL (rs487174 - PRDX1). However, it remains unclear which introns or exons are associated with the genetic variant as sQTL. Specifically, it is important to clarify whether it is the first intron of PRDX1 because the calculation method of promoter activity employed in this paper seems to be based on junction read counts on the first intron of the transcript, in which case inclusion ratio of the first intron would be almost the same thing as promoter activity. In this case, the comparison of puQTL and sQTL here would be just looking at the same thing. It would be beneficial for the authors to include the genomic coordinates and the number of the intron.

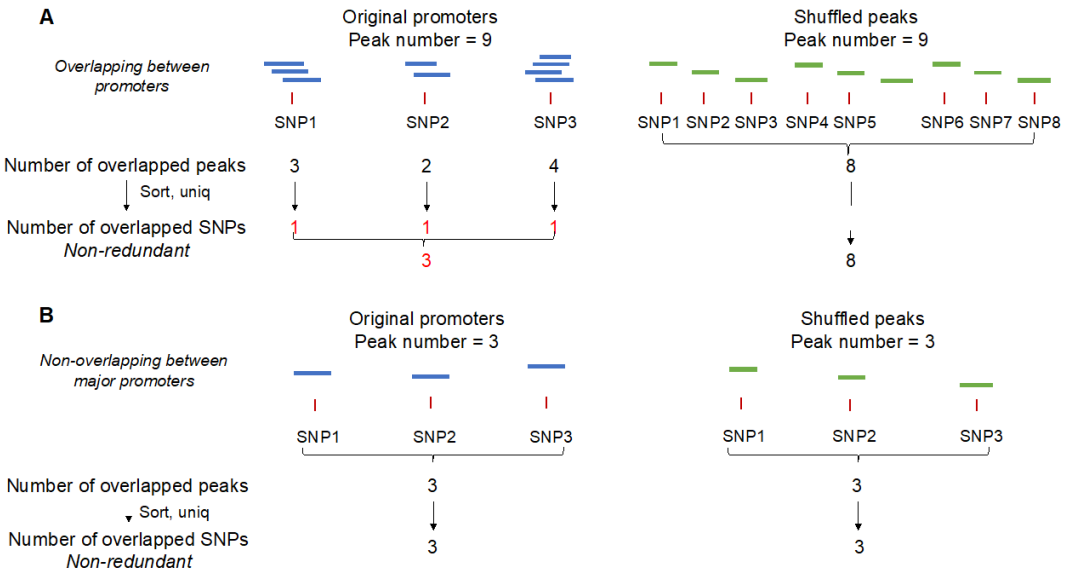
RESPONSE: Thank you for the insightful comments. We agree with the reviewer's suggestion regarding the necessity of clearly specifying the introns or exons associated with the sQTL. To address this concern, we meticulously examined the trait coordinates of the sQTL and have included the corresponding findings in **Figure S9**. Specifically, we have provided several illustrative examples to elucidate our findings, such as the association of rs487174 with the first intron of PRDX1 as a sQTL, the association of rs7170818 with the second intron of CHSY1 as a sQTL, and the association of rs4820023 and its adjacent four sQTLs with various exons and introns of the DUSP18 gene. It is important to note that the majority of sQTLs presented in the manuscript are not associated with the first intron of the gene. (**Page 19, Lines 436-437**)

9. Regarding comment #8, it is important for the authors to provide information about the source of the sQTL data and explain how the mapping of those data was conducted. The method of sQTL mapping should be clearly described.

RESPONSE: Thank you for the comments. Our study incorporates sQTL data obtained from QTLbase (Huang D, *et al., Nucleic Acids Res.,* 2023), a comprehensive database providing QTL summary statistics for a wide range of human molecular traits in over 95 tissue/cell types. The original sQTL data was obtained from GTEx v8 (Kim-Hellmuth S, *et al., Science*, 2020), which shares the same data resource as our study. (**Page 19, Lines 425-428**)

10. In Figure 3J and K, the authors demonstrate a significant depletion of eQTLs in promoters. Is this consistent with previous eQTL studies? Please cite some eQTL papers and compare the results.

RESPONSE: We would like to express our appreciation for the insightful feedback. Upon re-evaluating our results, we found a significant issue regarding the presence of numerous overlaps between adjacent promoters. Specifically, when overlapping SNPs with promoters, we noticed a high number of overlaps between adjacent promoters. After overlapping the SNP with the promoters, we counted the number of SNPs that had overlaps with promoters. This count is for a non-redundant SNP set. For instance, if a SNP is overlapped with three promoters, after applying the sort and uniq command, the SNP will only be counted once (**Response Figure 3A**). In contrast, the shuffled set, with the same peak number as the original promoters, generated by randomly shuffling the peaks across the entire chromosome, did not exhibit overlaps between shuffled peaks. As a result, the count of SNPs overlapped with promoter regions in the shuffled set was not affected by the sort and uniq command. Consequently, we found that the number of overlapped SNPs in the real dataset was underestimated, leading to the observed depletion of eQTLs in promoters. This same issue was also present in the 3'UTR and 5'UTR regions.



Response Figure 3. Framework of permutation test in last submission (A) and this version (B).

To address this challenge, we have made significant revisions to the manuscript. Specifically, for each gene, we filtered the transcripts by retaining only those with a major promoter. This approach ensures that only one set of promoter, 3'UTR, and 5'UTR is retained for each gene, effectively mitigating the impact of overlaps between adjacent regions (**Response Figure 3B**). In our revised results, we found that both puQTLs and eQTLs were significantly enriched in promoters, 3'UTRs, and 5'UTRs (see **Figures 3I-3K and S4C-S4D**). We have thoroughly revised the Results and Methods sections to reflect these major changes. (**Page 13, Lines 287-289; Page 29, Lines 670-674**)

11. In Figure S8D, the authors demonstrate that rs7738430 is linked to the activity of Promoter 68943 with a high beta value. However, the relationship does not appear to be linear, possibly because of the very low activity level in individuals with the aa genotype. Please verify the authenticity of this association and modify the figure to enhance readers' intuitive understanding.

RESPONSE: We express our gratitude to the reviewer for highlighting this issue. Upon thorough examination, we have confirmed that the variance of promoter activity falls below the

threshold in our updated datasets. As a result, we have replaced it with a new illustration in **Figure S11**. (**Page 21, Lines 470-474**)

Minor comments:

1. Related to major comment #4, please add a few sentences about how the promoter activity was estimated in this study.

RESPONSE: Thank you for the valuable comment. In response, we have incorporated a new paragraph within the Methods section to provide a detailed description of the approach utilized for assessing both absolute and relative promoter activity. This inclusion is intended to enhance the clarity and comprehensiveness of our manuscript for the benefit of our readership. (**Page 25, Lines 572-585**)

2. Lines 81-87: I don't understand why these last two out of the three mentioned points are related to concerns in the definition of promoter usage. The structure of the sentence should be reorganized.

RESPONSE: Thank you for the constructive suggestion, and we have made revisions accordingly. The third point has been removed, and the second point has been elaborated with additional illustrative examples to enhance clarity and depth of explanation. (**Page 5, Lines 85-94**)

3. Fig. 1E: Please add an asterisk next to "Fraction of ~~".

RESPONSE: Thank you for the meticulous review. In response to the reviewer's feedback, an asterisk has been incorporated into the figure. Furthermore, an increased number of promoters were found to harbor more than four puQTLs, coinciding with the expansion of the datasets. Consequently, we have denoted the tissues with an asterisk when the proportion of the group with more than 4 puQTLs exceeds 50%. (**Page 7, Lines 141-143**)

4. Please cite the original paper of Bedtools at line 597 on page 26.

RESPONSE: We appreciate the reviewer's reminder. The citation has been added to the text. (**Page 29, Lines 667-669**)

5. Fig. 5K: Please increase the size of the circles in the boxes.

RESPONSE: Thank you for the suggestion. The new figure has been updated as **Figure 5L**. We have adjusted the size of the circles in the figure and included additional examples in **Figure S8H** related to this issue. (**Page 19, Lines 431-435**)

6. Fig. 5C-D: How about prmtr.21008

RESPONSE: We addressed this in our response to Comment #7 from Reviewer #1.

7. Fig. 3K: It appears that the P-value is incorrect. Please correct it.

RESPONSE: We appreciate the reviewer for pointing out this issue. It has been rectified. Further elaboration has been included in the reply to Comment #10 from Reviewer #1.

Overall, this manuscript has great potential, but further revisions are needed to address the aforementioned points. With these improvements, the study will make a significant contribution to the scientific community.

RESPONSE: We express our sincere appreciation to the reviewer for recognizing the importance of our research. The constructive feedback provided has been instrumental in refining the biological insights presented in our manuscript. We have made revisions to address the comments.

Reviewer #2 (Remarks to the Author):

This paper describes a new analysis of the GTEX dataset to identify promoter usage QTLs (puQTLs), a class of gene regulation that is distinct from previous reports investigating eQTLs or promotor activity in the GTEX dataset. The introduction is clearly written, and the authors have clearly justified why this additional layer of regulation is worth investigating. The authors identify puQTLs and perform multiple analyses to characterize the properties of puQTLs, how they compare to eQTLs and paQTLs, and identity potential ways in which puQTLs expand upon previous QTL discoveries in this well studied dataset. The paper and analysis represent a significant body of work. However, the methods in several places are unclear, and several analytical points need to be clarified for the full significance of the findings to be interpretable.

RESPONSE: We would like to extend our appreciation for the valuable feedback and insightful comments, recognizing the potential enhancements in the manuscript's comprehensibility. In response to the comments, we have further revised and expounded upon the Methods section to enhance its clarity. Furthermore, in direct response to the suggestions, Response Figures 4-7 and Response Table 1 have been incorporated. Additionally, a data portal has been implemented to facilitate the sharing of summary statistics related to puQTLs, in accordance with the provide guidance.

Additionally, in accordance with the suggestions of Reviewer #3, the GTEx RNA-seq dataset has been updated from version 6 to version 8, encompassing a substantial increase in samples (from 6,674 to 15,201) and donors (from 503 to 838) across various tissues (**Response Figure 1**). Additionally, the transition to the GRCh38 human genome reference and GENCODE v26 genome annotation file aligns with the version utilized in GTEx v8. This expansion in sample size enhances the statistical power for mapping puQTLs across 49 tissues. It is important to acknowledge that due to the increased sample size and updated genome annotation, certain details such as promoter ID, effect size values, and false discovery rate (FDR) may have been revised, leading to corresponding changes in the figures. Nevertheless, the overarching genome-wide conclusions remain unchanged and may have

been strengthened by the heightened statistical power of the model. (**Page 3, Lines 35-37; Page 7, Lines 131-135**)

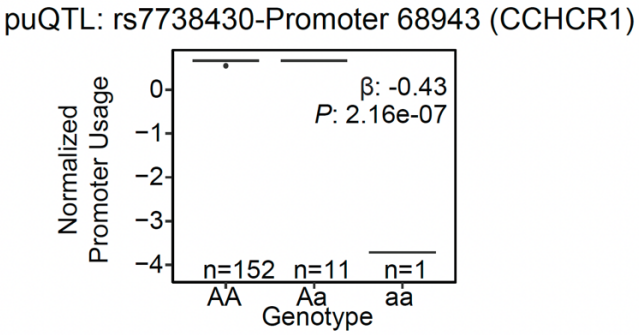
Major Points

1. The authors describe several filtering steps prior to analyses, including removing promoters with missing value exceeding 50% per tissue and 80% per individual. As the final value used is a proportion, ie the usage of the major promotor compared to other promotors, it is unclear why these filters were needed, and particularly if removing them could inflate the final proportion of the major promoter in individuals where minor promotors were filtered. Could the authors please explain their justification for selecting these missingness thresholds, and detail what impact they have on the final results?

RESPONSE: We appreciate the reviewer for highlighting the crucial issue. We have clarified the filtering pipeline to make it clearer for the reader to follow. Prior to the filtering, the relative promoter activity has been calculated, and it will not be affected by the filtering of minor promoters or individuals. In the revised version, we implemented five steps to prepare the promoter usage matrix for each tissue. To ensure data quality, we defined a missing value if the promoter usage is absent in a promoter or an individual. First, we selected the major promoter with the highest average relative promoter usage across all samples for subsequent analysis. Then, for each promoter, we filtered out promoters with a missing value in more than 50% of individuals. Next, for each individual, we excluded those with missing values in more than 80% of promoters. We also filtered out promoters with a standard deviation equal to 0 across all samples. Finally, the resulting promoter usage values were scaled and normalized using quantiles for each tissue. In the revised filtering steps, the major promoter was identified prior to the removal of minor promoters to prevent an overestimation of the relative proportion of major promoters.

Setting missingness thresholds is crucial for mitigating biases introduced by undetectable expression levels. For instance, the expression level of promoter 68943 is low in most samples, resulting in nearly zero variance of promoter usage after normalization. Consequently, even a

single sample in aa could significantly impact the statistical outcome (**Response Figure 4**), necessitating the removal of such examples through the filtering pipeline. Additionally, the filtering cutoff is widely employed in related research, such as eQTL analysis in GTEx v8 (https://gtexportal.org/home/methods) (Kim-Hellmuth S, *et al., Science*, 2020), m6A QTLs (Xiong X, *et al., Nat Genet.,* 2021), and 3'a QTLs (Li L, *et al., Nat Genet.,* 2021). In the context of 3' QTL analysis, it is important to highlight that the PDUI value serves as a measure of the percentage of distal poly(A) site usage index. This metric bears resemblance to the proportional promoter usage that we have employed in our investigation. We have modified the Methods section accordingly. (**Page 26, Lines 592-600**)



Response Figure 4. Examples demonstrating the essential of filtering steps.

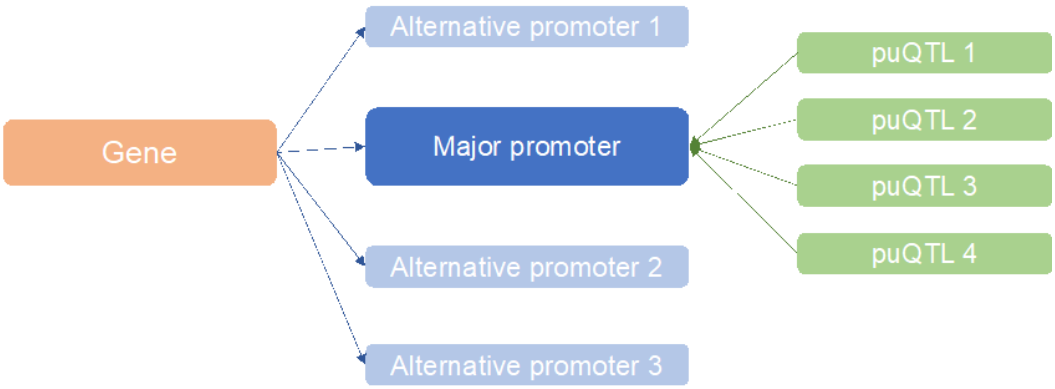
2. The reporting of the number of puQTLs is unclear. This is reported as 303,025 puQTLs across 48 human tissues. How many unique loci does this represent?

RESPONSE: We express our gratitude to the reviewer for the valuable comment. The distribution of puQTLs in various tissues is visually depicted in **Figure 1A** and quantified in **Table S1**. In our previous submission, we reported 303,025 distinct puQTL loci across 48 tissues, representing a non-repetitive set. Subsequent to the expansion of our dataset, the count of unique puQTL loci has been revised to 602,009. (**Page 7, Lines 131-133**)

3. The text states that only the most commonly used promoter was selected for further analysis and further that "Specifically, puQTL focuses on genes, assigning a single promoter usage value to each gene and correlating it with genotypes... Unlike paQTL, which considers all

promoters, puQTL identifies the most frequently utilized promoter to elucidate the correlation between genetic variations and promoter usage." However, in the text several mentions are made of genes with multiple puQTLs. It is unclear from the text how multiple puQTLs per gene have been identified, or what they represent. I would assume multiple puQTLs per gene were a result of conditional analyses to identify additional independent SNPs regulating the usage of the most frequent promoter? This is not detailed in the methods and text. The authors should explicitly explain how they defined multiple puQTLs per gene and add this to the methods.

RESPONSE: Thank you for the insightful comment. In the context of a specific gene, it's important to note that only one promoter can be designated as the major promoter. However, it's worth noting that while there's only one major promoter for a given gene, there can be multiple puQTLs independently associated with the usage of this major promoter. For example, in the case of gene *DUSP18*, the promoter 49007 serves as the major promoter, and the study identified five puQTLs significantly linked to the usage of this specific promoter (**Figure S7**). We have drawn a model to elucidate the regulatory relationships between multiple puQTLs and the major promoter (**Response Figure 5**). Additionally, we have clarified the relationships between multiple puQTLs and the major promoter in the Results and Methods section. (**Pages 18-19, Lines 407-420; Page 25, Lines 583-585**)



Response Figure 5. Regulatory patterns between multiple puQTLs and the major promoter.

4. The authors report the overlap betewen puQTL -gene pairs and paQTL pairs, and later with eQTLs. How were these overlaps calculated? Was only the lead SNP used for each dataset?

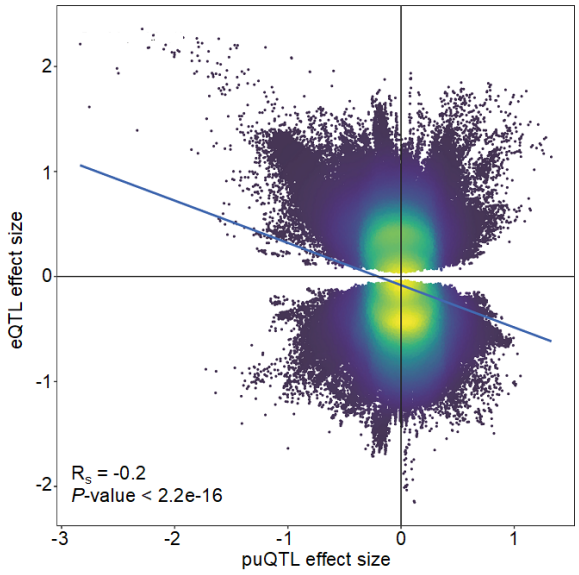
Or were linked SNPs included as well? If so was there an LD threshold applied between the lead and linked SNPs in order to identify true overlaps? A description of the methods used to calculate the overlaps should be added to the methods.

RESPONSE: Thank you for bringing this to our attention. Our analysis focused solely on the lead SNP for the overlapping analysis. When comparing puQTL-gene pairs and paQTL-gene pairs, the overlap was assessed at the SNP-gene pair level (**Figure 2B**). In contrast, the overlap between puQTL and eQTL was evaluated at the gene level, where we compared the genes associated with puQTL and paQTL to determine the number of overlapping genes (**Figure 3A**). Additionally, for genes associated with both puQTL and eQTL, we calculated the LD r^2 between lead eQTLs and puQTLs (**Figure 3D**). We have included a new paragraph detailing the methodology and dataset used for the overlap analysis in the Methods section. (**Page 28, Lines 654-660**)

5. In Figure 3F and 3G, the authors have split the dataset into directionally concordant subsets and then performed correlation analysis. This massively inflates the correlation and is not meaningful. Could the authors please display and report the overall correlation between eQTL effect size and puQTL size, including all examined pairs?

RESPONSE: We extend our gratitude to the reviewer for the valuable comment. It is worth noting that the effect size of puQTL demonstrates a statistically significant negative correlation with the effect size of eQTL (R_S=-0.2, *P*-value < $2.2e^{-16}$) (**Response Figure 6**). Notably, the dataset was not partitioned into directionally concordant subsets, as evidenced by the presence of numerous data points located in the second and fourth quadrants displaying opposing directions (**Figure 3G**). By categorizing the genes into subsets based on their positive and negative correlations, we were able to distinguish two distinct groups of genes that were consistently or inversely regulated by puQTLs and eQTLs. This approach allowed for the identification of specific gene regulatory patterns associated with the different types of QTLs. Furthermore, previous studies have also utilized the correlation of effect size for directionally concordant subset (Liang D, *et al., Nat Neurosci.,* 2021). (**Page 12, Lines 258-**

264)



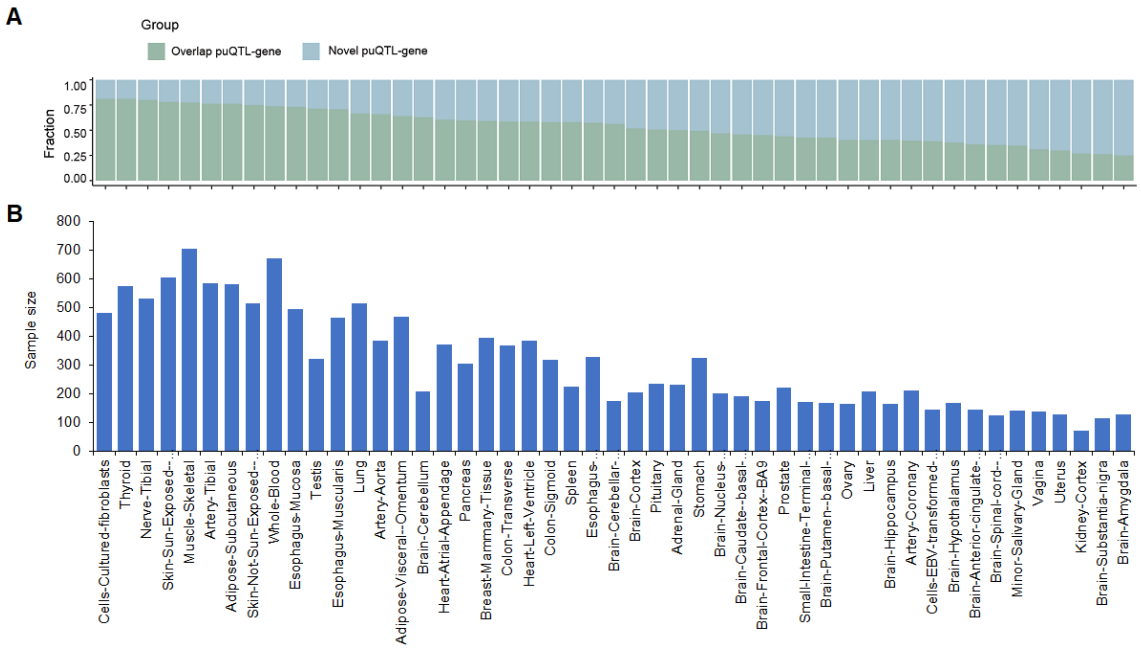
Response Figure 6. Overall correlation between effect size of puQTL and eQTL.

6. The authors report an overlap between the percent of genes with reported eQTLs and reported puQTLs. As the number of genes with an eQTL in the GTEX dataset (and eQTL datasets in general) is heavily dependent on sample size, surely the overlap here is mostly driven by the different sample sizes across tissues and subsequent power to detect QTLs rather than on any underlying molecular explanation?

RESPONSE: We extend our gratitude to the reviewer for the valuable insights regarding the comparison of overlapping proportions between puQTL and eQTL across various tissues. Upon further investigation, we observed a strong correlation between the sample size of each tissue and the extent of overlap, in line with the reviewer's expectations (**Response Figure 7**). Our findings indicate that larger sample sizes result in increased power to detect both eQTLs and puQTLs, consequently leading to a greater overlap between the two types of QTLs. The sample size chart has been incorporated into **Figure S4A**, with accompanying explanatory text to facilitate the interpretation of the data.

Additionally, inspired by the reviewer's feedback, we have also improved the results in **Figure S5A**. Furthermore, we consistently observed that as the sample size increased, a greater number of puQTLs were found to be specific to a particular tissue, indicating an

increase in the model's ability to detect puQTLs and resulting in the identification of more tissue-specific puQTLs. (**Page 12, Lines 264-270; Page 14, Lines 307-312**)



Response Figure 7. Overlapping between puQTL-genes and eQTL-gene across different tissues (A). The sample size of each corresponding tissue (B).

7. The authors state “Remarkably, puQTLs displayed a significantly stronger correlation with GWAS traits in comparison to paQTLs and eQTLs. “ But they do not clarify what evidence this statement is based on. Could the authors please clarify what evidence supports this strong statement? There are several potential relevant figures presented in the supplemental figure, but these are not clear in the text and I was not convinced that this statement is supported by evidence.

RESPONSE: We thank the reviewer for the insightful comments, and we have revised the manuscript accordingly. Our conclusion is drawn from the analysis of data presented in **Figures 6A, 6B, and S11A**. We observed that the difference between observed p-values and expected p-values is more significant for puQTLs compared to paQTLs and eQTLs. As our conclusion is based on multiple tissues, we have adjusted the statement to avoid any overgeneralization. In the revised manuscript, we specifically highlight the functional interpretation of puQTLs, stating that "puQTLs demonstrate a strong correlation with GWAS

traits, indicating their functional significance." (**Page 20, Lines 454-455**)

8. In the colocalization methods, the authors state: “To conduct the colocalization analysis of puQTL-promoter pairs, we first obtained the significant SNPs from each GWAS trait (P-value < 5 x 10-8) and subsequently compared them with the nominal puQTLs (FDR < 0.05). “ I do not think this approach is valid. The method they used, Coloc requires full summary statistics from both traits. Inputting only the GWAS significant hits and nominal puQTLs into COLOC will massively inflate the number of co-localizations. This analysis should be rerun with full summary statistics of both traits.

RESPONSE: We express our gratitude to the reviewer for the constructive comments. In response to comment, we have made significant improvements to our methodology for colocalization analysis. Specifically, in the revised manuscript, we have expanded our inclusion criteria to encompass all puQTLs and a broader range of GWAS traits-associated SNPs ($P < 1 \times 10^{-5}$), thereby relaxing the previously imposed cutoff threshold. It is noteworthy that many previous studies on colocalization analysis of QTLs have employed cutoff thresholds for either QTLs or GWAS traits (**Response Table 1**).

Response Table 1. Cutoff for selecting QTLs or GWAS traits in previous publications.

Type of QTL	Cutoff for QTLs	Cutoff for GWAS traits	Publication
m6A QTL	N.A.	$P < 1 \times 10^{-4}$	Xiong X, *et al., Nat Genet.,* 2021
Trans eQTL	$P < 1 \times 10^{-5}$	N.A.	Kim-Hellmuth S, *et al., Science*, 2020
3'a QTL	N.A.	$P < 5 \times 10^{-8}$	Li L, *et al., Nat Genet.,* 2021
haQTL	Empirical P < 0.05	$P < 1 \times 10^{-5}$	Hou L, *et al., Nat Genet.,* 2023
cQTL	FDR < 0.05	N.A.	Bryois J, *et al., Nat Commun.,* 2018
caQTL	FDR < 0.05	$P < 5 \times 10^{-8}$	Liang D, *et al., Nat Neurosci.,* 2021
tuQTL	Within 100kb	$P < 1 \times 10^{-5}$	Alasoo K, *et al., eLife,* 2019
puQTL	None, using full set	$P < 1 \times 10^{-5}$	This study

N.A. means not specifically mentioned in the study.

For instance, in the GTEx project, only trans-eQTLs with $P < 1 \times 10^{-5}$ were utilized for colocalization analysis (Kim-Hellmuth S, *et al., Science*, 2020). Similarly, in the investigation of chromatin QTLs (cQTLs), only cQTLs with an FDR < 0.05 were considered for colocalization (Bryois J, *et al., Nat Commun.,* 2018). Other studies have also employed specific cutoff thresholds for GWAS trait SNPs, such as setting $P < 1 \times 10^{-4}$ for m6A QTLs (Xiong X, *et al., Nat Genet.,* 2021) and $P < 5 \times 10^{-8}$ for 3'a QTLs (Li L, *et al., Nat Genet.,* 2021). Furthermore, studies focusing on histone acetylation QTLs (haQTLs), chromatin accessibility QTLs (caQTLs) and transcript usage (tuQTLs) have implemented cutoff criteria for both QTL sets and GWAS traits (Hou L, *et al., Nat Genet.,* 2023; Liang D, *et al., Nat Neurosci.,* 2021; Alasoo K, *et al., eLife,* 2019). In conclusion, the established cutoff criteria for GWAS traits were uniformly implemented and widely endorsed within the discipline. (**Page 32, Lines 756-759**)

9. The authors do not include what methods/parameters they used for colocalization of the paQTLs or the eQTLs. This should be added.

RESPONSE: Thank you for the comments. We have revised the Methods section for colocalization analysis by adding detailed parameters (**Page 33, Lines 767-773**).

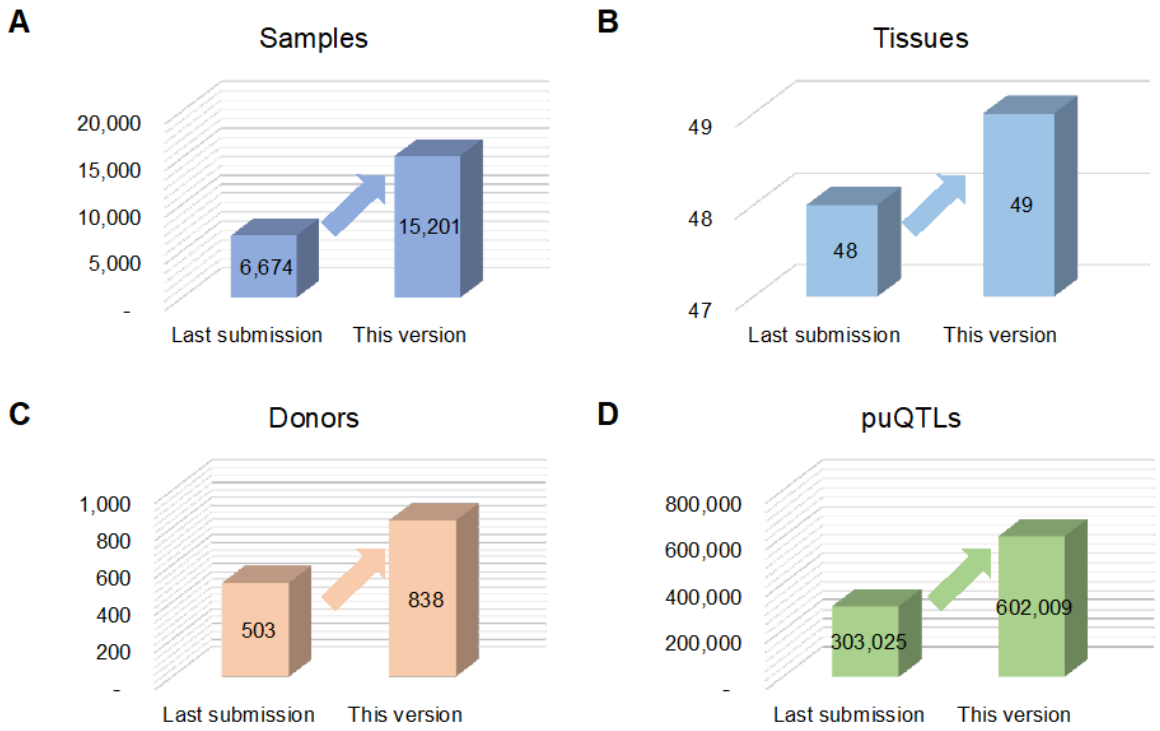
10. The Data Availability statement does not detail how the results of this study – the puQTL summary statistics will be made available.

RESPONSE: Thank you for the constructive suggestions. In response, we have developed a user-friendly data portal (http://bioailab.com:3838/GTEX_puQTL/) to facilitate the browsing and retrieval of the puQTL summary statistics outlined in our study. This portal enables users to explore puQTLs by selecting specific tissue types and conducting searches using SNP IDs or gene symbols. Moreover, we have included a download link for immediate access to the data. We believe that this resource will greatly benefit the scientific community by providing access to puQTL data for further investigation and analysis. We have updated the Data Availability section of the manuscript to include the website link. (**Page 34, Line 799-804**)

Reviewer #3 (Remarks to the Author):

In this paper, Yuan et al present an analysis of promoter usage QTLs (puQTL) in a subset of GTEx samples. While the paper presents some interesting results about the properties of puQTLs, its utility as a resource is strongly limited by using only a subset of GTEx samples, reliance on outdated gene annotations (GENCODE v19, released in 2013), and a complete lack of publicly available summary statistics. Furthermore, the introduction of the paper overclaims the novelty of the work by misrepresenting some of the previous work.

RESPONSE: We express our gratitude towards the comments. In response to the feedback, several significant revisions have been implemented in the manuscript. Primarily, the GTEx dataset has been updated to version 8, resulting in a substantial augmentation in the sample size (from 6,674 to 15,201), donor count (from 503 to 838), and puQTLs (from 303,025 to 602,009) (**Response Figure 1**). Furthermore, the genome annotation and reference genome have been updated to ensure alignment with the GTEx v8 (utilizing GENCODE v26 for genome annotation and GRCh38 for the reference genome). Additionally, a dedicated online data portal has been implemented to facilitate seamless access to the summary statistics of puQTLs (http://bioailab.com:3838/GTEX_puQTL/), providing researchers with an avenue to explore and leverage this valuable resource for further analysis. These revisions not only address the concerns but also significantly enhance the accessibility and applicability of the puQTL data within the research community. (**Page 3, Lines 35-37; Page 7, Lines 131-135**)



Response Figure 1. Comparisons of the number of samples **(A)**, tissues **(B)**, donors **(C)** and puQTLs **(D)** between last submission and this revised version.

Major concerns

1. On lines 73-78, the authors make two claims: 1. “Prior investigations commonly evaluate promoter activity in a restricted set of tissues. However, there is a notable absence of a comprehensive study that examines promoter usage across a wide range of tissues. 2. “Besides, the strategy used in these studies is to measure promoter activity for each isoform by using the absolute expression level of promoter regions (19,20).”

Both of these claims are demonstrably false. First, these claims ignore the work by Alasoo, et al 2019 (1), which quantified the usage (not the activity) of each promoter relative to all other promoters of the same gene, thus exactly following the puQTL naming convention in this paper. While there are methodological differences in how promoter usage is quantified in this paper versus how it was quantified by Alasoo et al, it is not obvious that one of the approaches is superior to the other and no such comparison is presented in the paper. Secondly, the txrevise method developed by Alasoo et al to quantify promoter usage (along with internal exon usage and 3’end usage) has since been applied to all RNA-seq datasets present in the eQTL Catalogue (https://www.ebi.ac.uk/eqtl/) (2). These summary statistics have been publicly

available since 2020 and have been continuously updated, most recently in April 2023 (https://www.ebi.ac.uk/eqtl/Release_notes/) (3).

Thus it is false to claim that previous studies have not quantified promoter usage (as opposed to promoter activity) and that this analysis has not been extended to a large number of tissues and cell types. The authors should rewrite the introduction of their paper to accurately reflect the published literature.

RESPONSE: We would like to extend our appreciation to the reviewers for drawing attention to several published studies on puQTLs. Upon careful review of these publications, we have realized that we have inadvertently failed to consider these studies, which have already defined the promoter usage and identified a number of puQTLs. In the revised manuscript, we have appropriately referenced these studies in the Introduction section to fully acknowledge their influence on the field.

Although previous studies have defined puQTLs, our research diverges from the current literature in this area. For instance, the study published in *eLife* focused on mapping puQTLs within stimulated macrophages, identifying 3,500 puQTLs related to promoter usage, splicing, and 3' end usage. Conversely, a recent work in *Nat Genet.* 2021 primarily investigated eQTL and sQTL, neglecting puQTLs. Additionally, no information about puQTLs was found on the specified website under data access and release notes. Another study in *PLoS Genet.* 2023 reported only 7 puQTLs colocalized with GWAS traits without providing the total number of puQTLs. In summary, while the earliest paper in *eLife* focused primarily on puQTLs within a single cell type, the other two studies encompassed multiple tissues or cell types without emphasizing puQTLs.

Our study employed a multifaceted approach to elucidate the tissue-specific characteristics of puQTLs and their enrichment within actively transcribed regulatory regions. Additionally, integration of ChIP-seq data related to epigenetic markers and HiChIP data was employed to corroborate the interactions between puQTLs and promoters. Importantly, our study revealed the colocalization of puQTLs with GWAS signals for complex traits/diseases and their contribution to heritability. Overall, our study not only thoroughly delineates puQTLs across 49 distinct tissues, but also posits puQTLs as a new category of molecular quantitative

loci across various strata. As a result, our study provides a distinctive perspective that set it apart from previous publications in the field. (**Page 5, Lines 77-81**)

2. It is unclear how many RNA-seq samples were included in the analysis. The abstract mentions 503 individuals, which is significantly less than 838 donors present in the GTEx version 8 release (published in 202) (4). Furthermore, the GTEx v8 release contained data from 15,201 samples, but the number of samples analysed in this study is unclear as it is not mentioned anywhere in the paper. The method section only refers back to the paper by Demircioglu et al (5), which only analysed 6,674 GTEx samples (44% of GTEx v8). Is 6,674 the sample size used in this study? If yes, then the authors should clearly state this early in the paper. They should also be explicit that they have used less than half of the data available from GTEx. Moreover, the latest release of the eQTL Catalogue contains uniformly processed puQTLs for 25,724 samples from 109 datasets, which is significantly larger than the current study.

RESPONSE: We would like to thank the reviewer for the valuable feedback. It has come to our attention that the previous version of the manuscript did not sufficiently elucidate the sample size of the GTEx RNA-seq dataset used. In the previous version, we indicated that the estimation of promoter activity was based on the study by Demircioglu *et al*., encomprising 6,674 samples, as detailed in the Methods section. In response to the reviewer's suggestions, we have updated our utilization of the GTEx dataset from version 6 to version 8. Specifically, we obtained the RNA-seq junction read files comprising 15,201 samples within the GTEx v8 project and subsequently employed the *proActiv* tool to evaluate both the absolute and relative promoter activities. Consequently, the enhancement of our study's robustness is evident through utilization of the GTEx v8 dataset, which has augmented the sample size from 6,674 to 15,201 (**Response Figure 1**). To ensure transparency and clarity, we have revised the manuscript to reflect the updated sample size. We extend our appreciation to the reviewer for the valuable comments in this regard. (**Page 3, Lines 35-37**)

3. The methods section currently does not provide any information on how the promoter

activity values were computed from the GTEx RNA-seq data. Please add information about which software was used and with which parameters. Also, it would be important to justify why you decided to use transcript annotations from GENCODE v19 (released over 10 years ago) when much newer annotations are available. Finally, GENCODE v19 was based on GRCh37 coordinates while the WGS genotype data that you used was based on the GRCh38 coordinates. How did you reconcile this difference? Did you use liftOver or some other tool to convert the coordinates from one version of the reference genome to the other? If yes, then this should also be explicitly mentioned in the methods.

RESPONSE: Thank you for the constructive comments. We have addressed the comment about the lack of explanation for how we calculated promoter activity from RNA-seq data by updating the Methods section. We employed a tool called *proActiv* to measure both the absolute and relative promoter activity by counting junction reads. The junction read count files were obtained from the GTEx portal as instructed, and we collected 15,201 junction read files in the SJ.out.tab format. The R package, *proActiv* developed by Demircioglu *et al.*, was developed to estimate promoter activity using a set of unique junction reads (Demircioglu D, *et al., Cell*, 2019). The performance of *proActiv* in promoter activity estimation has been benchmarked using H3K4me3 histone data, CAGE data and Iso-seq data (Huang KK, *et al., Genome Biol.,* 2021).

In response to the comment regarding the genome annotation and reference genome versions, it is pertinent to note that our previous manuscript utilized the GTEx V6 RNA-seq dataset (GRCh37). Consequently, we converted the genotype data to align with the GRCh37 reference genome using liftOver. However, it is imperative to highlight that these concerns have been rectified in the revised manuscript. Specifically, the GTEx RNA-seq dataset has been updated from version 6 to version 8. Furthermore, the junction read files for promoter activity estimation, comprising 15,210 samples, were obtained from the GTEx portal in SJ.out.tab format (https://gen3.theanvil.io/explorer). Notably, the revised manuscript updated genome annotation and reference genome to ensure compatibility with GTEx v8 dataset, leveraing GENCODE v26 for genome annotation and GRCh38 for the reference genome.

The constructive feedback from the reviewer is greatly appreciated. The revisions made

to the manuscript are intended to improve its lucidity and accuracy, thereby promoting a more thorough understanding and recognition of the results. (**Page 7, Lines 128-131; Page 25, Lines 572-585**)

4. The title of the paper starts with “An atlas …” which implies that it describes a resource that could be used by other researchers. Is this true? If yes, then it would be important to make the summary statistics from the analysis publicly available. Currently, the data availability statement does not mention summary statistics at all.

RESPONSE: We express our sincere appreciation for the constructive feedback provided by the reviewer, which has the potential to significantly augment the impact of our study. We have developed a user-friendly data portal (http://bioailab.com:3838/GTEX_puQTL/) to facilitate the browsing and retrieval of the puQTL summary statistics detailed in our study. This portal allows users to explore puQTL by selecting specific tissue types and conducting searches using SNP IDs or gene symbols. Additionally, we have incorporated a download link to provide immediate access to the summary statistics data. We believe that this resource will substantially benefit the research community by granting access to puQTL data for further investigation and analysis. Furthermore, we have updated the Data Availability statement of the manuscript to include the website link. (**Page 34, Lines 799-804**)

References

1. K. Alasoo, J. Rodrigues, J. Danesh, D. F. Freitag, D. S. Paul, D. J. Gaffney, Genetic effects on promoter usage are highly context-specific and contribute to complex traits. Elife. 8 (2019), doi:10.7554/eLife.41673.

2. N. Kerimov, J. D. Hayhurst, K. Peikova, J. R. Manning, P. Walter, L. Kolberg, M. Samoviča, M. P. Sakthivel, I. Kuzmin, S. J. Trevanion, T. Burdett, S. Jupp, H. Parkinson, I. Papatheodorou, A. D. Yates, D. R. Zerbino, K. Alasoo, A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat. Genet. 53, 1290–1299 (2021).

3. N. Kerimov, R. Tambets, J. D. Hayhurst, I. Rahu, P. Kolberg, U. Raudvere, I. Kuzmin, A. Chowdhary, A. Vija, H. J. Teras, M. Kanai, J. Ulirsch, M. Ryten, J. Hardy, S. Guelfi, D. Trabzuni,

S. Kim-Hellmuth, W. Rayner, H. Finucane, H. Peterson, A. Mosaku, H. Parkinson, K. Alasoo, eQTL Catalogue 2023: New datasets, X chromosome QTLs, and improved detection and visualisation of transcript-level QTLs. PLoS Genet. 19, e1010932 (2023).

4. The GTEx Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 369, 1318–1330 (2020).

5. D. Demircioğlu, E. Cukuroglu, M. Kindermans, T. Nandi, C. Calabrese, N. A. Fonseca, A. Kahles, K.-V. Lehmann, O. Stegle, A. Brazma, A. N. Brooks, G. Rätsch, P. Tan, J. Göke, A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. Cell. 178, 1465–1477.e17 (2019).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have properly addressed most of the original concerns in the first manuscript. However, there are still a few areas for improvement.

1. As for fine-mapping, what is the "Ratio of puQTL with causality"? Is this posterior inclusion probability (PIP)? If yes, it would be weird that >75% puQTL have causality effects. In general, only a few variants within a QTL set should be called "fine-mapped variants" with high PIP. Please clarify the way how the authors obtained this metric.
2. The authors should include the fraction of puQTL with high causality effects in epigenetically-supported puQTL-gene pairs. Do rs487174 show a high causality effect on Promoter 21197?
3. For the example of sQTL in the CSHY1 gene, the consistency between puQTL and sQTL does not represent interconnections between different layers of gene regulation; the intron associated with sQTL is used only in a transcript transcribed from Promoter 27819, in which case it is not a functional interplay between alternative splicing and promoter selection, but it is just looking at the same thing. The authors should tone down the claim in the manuscript.
4. Related to the previous comment, as an illustration of sQTL mapping (Figure S9), the authors should include the genomic coordinates of not only introns associated with sQTLs but also other introns with the same intron cluster. This sQTL analysis is based on the intron usage ratio quantified by leafCutter, in which case it is important to show associations between puQTL and all introns within the intron cluster to demonstrate if the consistency between puQTL and sQTL is just looking at almost the same thing.
5. As for Figure 5A, it would be better to show enrichment results for all examined tissues.

Reviewer #2 (Remarks to the Author):

The authors have answered several of my queries, however there are still major points that require clarification.

Major points
Point 1: I thank the authors for their clarification of the filtering procedure and have no further comments.

Point 2: I remain very confused with the authors tally of the number of loci discovered and their methodology for calculating multiple puQTLs for a single major promoter. There are several places in the text where authors use the terms 'puQTL loci' and 'genetic variant' interchangeably, suggesting they are mis-using these terms, and counting every genetic variant with an association below their thresholds as an individual puQTL, rather than controlling for linked SNPs within a locus. This has serious implications for all subsequent analyses.

For example, the response states: "Subsequent to the expansion of our dataset, the count of unique puQTL loci has been revised to 602,009". But in the revised abstract it states: "By constructing an atlas of human puQTLs across 4935 different tissues from 838 individuals, we have identified approximately 602,009 genetic variants associated with promoter usage."

The authors should clearly state how many independent loci these 602,009 genetic variants correspond to.

Point 3: Related to point 1 above, I am very worried that the authors are conflating discovery of multiple independent puQTLs with the presence of multiple SNPs in LD within a single loci/puQTL. This discrepancy must be resolved and corrected prior to any publication.

For example, the authors state in the response:
However, it's worth noting that while there's only one major promoter for a given gene, there can be multiple puQTLs independently associated with the usage of this major promoter. For example, in the case of gene DUSP18, the promoter 49007 serves as the major promoter, and the study identified five puQTLs significantly linked to the usage of this specific promoter (Figure S7). We have drawn a model to elucidate the regulatory relationships between multiple puQTLs and the major promoter (Response Figure 5). Additionally, we have clarified the relationships between multiple puQTLs and the major promoter in the Results and Methods section. (Pages 18-19, Lines 407-420; Page 25, Lines 583-585)

The referenced text from line 407 states:
In a similar vein, we identified a group of five adjacent SNPs in conjunction with promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase (DUSP) family, which serve as critical modulators subsequent to acute lung injury. We observed long-range H3K27ac mediated chromatin interactions between the SNPs and promoter 49007, and the promoter was bound by H3K27ac411 and H3K4me1 in lung tissue (Figure S7A). The puQTLs rs5749148 and rs5753260 displayed significant associations with the upregulation of promoter 49007, while puQTLs rs5753259, rs5749147, and rs4820023 were significantly linked to the downregulation of its usage (Figures S7B and S7C). Furthermore, the aforementioned five SNPs, characterized as potential paQTLs of promoter 49007, did not exhibit a significant effect on the activity of the promoter itself.

However, the example 'independent' SNPs rs5749148 and rs5753260 are in near complete LD (R2 = 0.996 in EUR), and in very high LD with the remaining three SNPs (R2= 1.0, 0.8 and 0.69). I cannot see how these five tightly linked SNPS could possibly represent five "multiple puQTLs independently associated with the usage of this major promoter." If a claim of multiple independent puQTLs for a single promotor are to be included in the manuscript the authors MUST perform analysis to demonstrate they are independent and not merely one signal that is being tagged by multiple linked SNPs. The authors should also clarify if subsequent analysis throughout the paper including all 602,009 genetic variants or if they have used a reduced set representing independent SNPs (ie one lead SNP per independent loci), and how many total independent puQTLs they identify.

Point 4: I thank the reviewer for their response to my original point. However, given their clarification, I think the following manuscript statement is misleading: However, as less than 16% of overlapped genes were associated with the same SNPs in both puQTLs and eQTLs, suggesting the presence of unique associations between these two types of QTLs and their respective targets (Figure3C)."Please expand the above results statement to either clearly state it only refers to the comparison on LEAD eQTL SNPs, and add a caveat explaining that the 16% overlap is likely a large underestimate as the lead SNP can vary considerably in regions with high LD.

Point 5 – I appreciate the additional analyses provided by the authors. I would suggest Response Figure 6 and the reported overall effect size correlation should be included in the manuscript supplemental material

Point 6 – I thank the authors for their clarification

Point 7 - I thank the authors for their clarification

Point 8 – I thank the authors for their comprehensive response, but I believe they have misunderstood

my query. I am uncertain if this is a simple mis-understanding of their description of their pipeline on my part or a methodological mistake, and I request the authors to clarify carefully.

The authors description of their coloc analyses has left me worried that they ran the coloc package on the subset of SNPs that met a significance threshold rather than including ALL SNPs in the genomic region that contains the significant loci. For example, the new methods text states, "To conduct the colocalization analysis of puQTL-promoter pairs, we first obtained the significant SNPs from each GWAS trait (P-value < 1 x 10-5) and subsequently compared them with the complete puQTL set, encompassing SNPs located within a 1Mb proximity of the interested promoter." This description implies that only SNPs which had been either p<10-5 in a GWAS trait or called as a puQTL were included in the co-localization.

While I appreciate the Response Table 1, the content lists the significance cut off values others studies have used for determining which loci to colocalize. These studies subsequently used ALL SNPs in the genomic regions that they co-localized. I would like the authors to be clear that in the actual co-localization between a puQTL loci and a GWAS they used ALL SNPs in the genomic region – not just the SNPs that passed a threshold. Trimming the input SNPS by a significance threshold will invalidate coloc.

I have pasted the instructions from the coloc website below:
First: select the SNPs to study. SNPs must * have summary data available in both studies * cover a single genomic region (perhaps defined by distance about some GWAS peak or delimited by recombination hotspots), and * represent a dense coverage of the region

Please do include all SNPs you have in the region: **do not trim by significance or MAF or other means.** Coloc enumerates all possible configureations of shared and non-shared causal variants and evaluates their relative support; it cannot do this if it does not have all the SNPs available.

Point 9: I thank the authors for their methods clarification

Point 10: I thank the authors for adding these data availability resources

Reviewer #3 (Remarks to the Author):

The authors have addressed all of my concerns about the content of the paper. However, I would still strongly recommend also publicly releasing the full promoter usage QTL summary statistics. This would allow other research to use the results from this paper in their analysis and would significantly increase the impact of the paper.

Point-by-point address of the comments raised by the reviewers for manuscript NCOMMS-23-45003A

We extend our heartfelt appreciation to the reviewers for their meticulous assessment of our revised manuscript and for providing valuable feedback. The profound insights have significantly improved the quality of our paper. Please find a detailed response to the reviewers' comments below.

It is important to note that the genotypes of SNPs have been reformatted in all figures, utilizing the reference and alternative allele in place of the traditional "AA", "Aa", and "aa" nomenclature throughout the manuscript. As a result, it is possible that the orientation of ß values may be altered accordingly. All the results regarding to this issue have been revised.

Reviewer #1 (Remarks to the Author):

The authors have properly addressed most of the original concerns in the first manuscript. However, there are still a few areas for improvement.

RESPONSE: We express our gratitude for the insightful and constructive comments provided by the reviewer. In response, additional analyses have been implemented in the revised manuscript. A detailed response addressing the reviewer's feedback is presented herein.

1. As for fine-mapping, what is the "Ratio of puQTL with causality"? Is this posterior inclusion probability (PIP)? If yes, it would be weird that >75% puQTL have causality effects. In general, only a few variants within a QTL set should be called "fine-mapped variants" with high PIP. Please clarify the way how the authors obtained this metric.

RESPONSE: We express our gratitude to the reviewer for highlighting this issue. It is important to note that the term "Ratio of puQTL with causality" is distinct from posterior inclusion probability (PIP). The primary objective of the analysis was to pinpoint puQTLs with causal relationships to each promoter using a fine-mapping approach. Given that promoters can have multiple puQTLs, some may exhibit causality while others may not. We calculated the ratio of puQTLs demonstrating causality by dividing the number of such puQTLs by the total number of puQTLs associated with a given promoter, resulting in the aforementioned ratio. In the initial submission, no PIP cutoff was established, leading to a ratio of puQTLs with causality exceeding 75%. In the

revised version, a PIP cutoff of 0.5 was implemented, resulting in a reduced ratio of puQTLs with causality falling below 25% (**Supplementary Fig. 13**). (**Page 21, Lines 475-478**)

2. The authors should include the fraction of puQTL with high causality effects in epigenetically-supported puQTL-gene pairs. Do rs487174 show a high causality effect on Promoter 21197?

RESPONSE: We thank the reviewer for the valuable feedback. In our analysis, we examined the overlap between puQTLs supported by HiChIP data and those exhibiting a high causality effect (PIP >= 0.5) in 22 different tissues, as illustrated in **Fig. 1f**. This allowed us to calculate the proportion of puQTLs that demonstrate both chromatin interaction and a strong causality effect, termed as the overlapped fraction of puQTLs. Our results reveal that about 40% of puQTLs with a high causality effect were corroborated by HiChIP data across the 22 tissues examined (see **Supplementary Fig. 13c**). While we observed chromatin interactions involving rs487174 and promoter 21197, rs487174 was not deemed to have a causal impact on promoter 21197, suggesting that integrating multi-epigenetic omics data could offer an alternative approach to pinpoint functional SNPs. (**Page 21, Lines 478-480**)

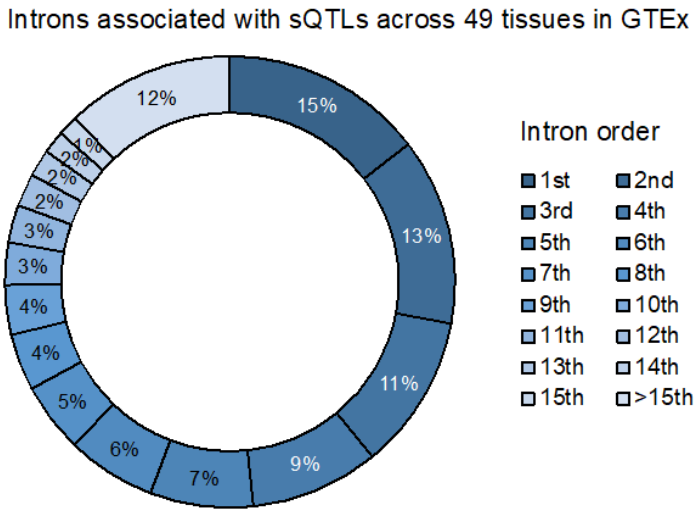
3. For the example of sQTL in the CSHY1 gene, the consistency between puQTL and sQTL does not represent interconnections between different layers of gene regulation; the intron associated with sQTL is used only in a transcript transcribed from Promoter 27819, in which case it is not a functional interplay between alternative splicing and promoter selection, but it is just looking at the same thing. The authors should tone down the claim in the manuscript.

RESPONSE: The constructive feedback provided by the reviewer is greatly appreciated. While our initial analysis demonstrated some consistency between puQTLs and sQTLs through specific examples, a genome-wide conclusion could not be drawn. As in response to comment #4, additional analyses were conducted to compare puQTL and sQTL on a genome-wide scale. The results of the genome-wide comparison revealed that approximately 20% of puQTLs were associated with the same introns as sQTLs, suggesting that puQTL and sQTL target distinct entities. We concur with the reviewer's assertion that drawing conclusions solely from a limited set of examples is inappropriate. In order to prevent any misinterpretation regarding the comparison between puQTLs and sQTLs, the relevant results have been omitted from the manuscript.

4. Related to the previous comment, as an illustration of sQTL mapping (Figure S9), the authors should include the genomic coordinates of not only introns associated with sQTLs but also other introns with the same intron cluster. This sQTL analysis is based on the intron usage ratio quantified by leafCutter, in which case it is important to show associations between puQTL and all introns within the intron cluster to demonstrate if the consistency between puQTL and sQTL is just looking at almost the same thing.

RESPONSE: We thank the reviewer for the insightful comments. To address the potential concern regarding that the consistency between puQTLs and sQTLs is just looking at the same thing, we have implemented a comprehensive comparison between puQTLs in our study and sQTLs identified in GTEx project.

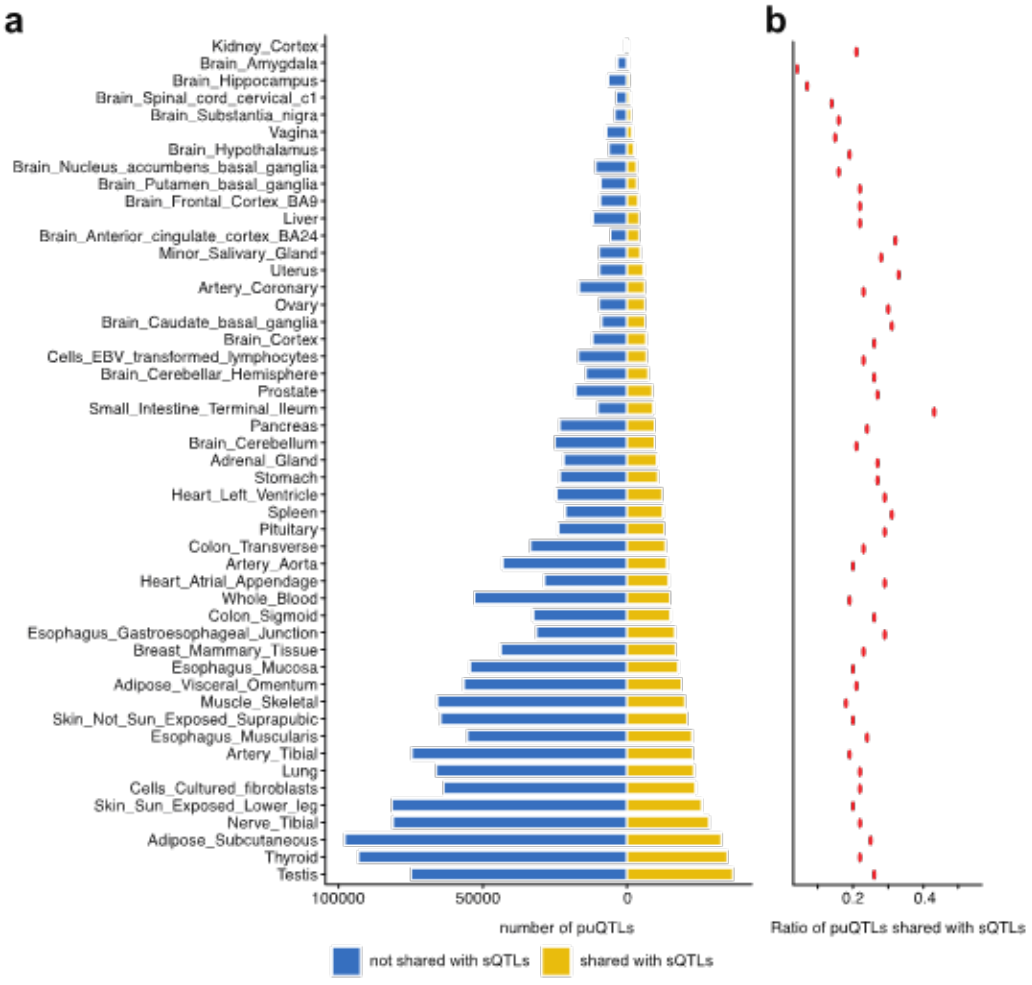
First, the quantification of promoter activity through the *proActiv* algorithms involves the utilization of the junction read count method, wherein only junction reads aligning to the first introns of the transcript set are tallied. It is crucial to highlight that the exclusion of internal promoters in the method is deliberate, as junction reads mapping to internal introns could potentially lead to ambiguous assignments to the promoters. Conversely, sQTL analysis focuses on intron usage, encompassing internal introns. Notably, our findings indicate that a mere 15% of sQTLs are linked to first introns, with a higher proportion (85%) associated with internal introns (see **Response Figure 8**). Thus, from a methodological standpoint, puQTL and sQTL estimations are conducted on distinct scales.



Response Figure 8. The order of introns associated with sQTLs across 49 tissues in GTEx.

Secondly, in accordance with the reviewer's feedback, all introns within the intron cluster

have been incorporated for the purpose of assessing the consistency between puQTL and sQTL. Our comparative analysis revealed that approximately 20% of puQTLs demonstrated significant association with the same intron as sQTLs across 49 distinct tissues (**Response Figure 9**), with notable instances involving rs7170818 and the gene *CHSY1*. While some puQTLs were found to affect the same intron of a gene as sQTLs, the bulk of puQTLs (with a median of 77%) exhibited effects on different targets or displayed opposite directional impacts on the same target. These results indicate distinct regulatory mechanisms governing puQTLs and sQTLs.



Response Figure 9. Comparison between sQTLs and puQTLs across 49 tissues in GTEx. (**a**) The number of puQTLs shared (yellow) and not shared with sQTLs (blue) across 49 tissues in GTEx. (**b**) The ratio of puQTLs shared with sQTLs in each tissue.

Thirdly, and perhaps most importantly, our study aims to differentiate between puQTLs and sQTLs in terms of their biological significance. The genetic variants identified as puQTLs are

linked to the transcription start site of a gene, representing a pre-transcriptional process. Conversely, the genetic variants identified as sQTLs are associated with alternative splicing of a transcript, indicative of post-transcriptional regulation that occurs subsequent to transcription. These distinct classifications of QTLs have unique definitions and calculation methodologies, resulting in disparate target sets and regulatory patterns. Therefore, it is imperative to recognize that puQTLs and sQTLs are not interchangeable, as each provides additional insights into the functional implications of genetic variants.

Due to the omission of three specific examples pertaining to the sQTL results in the manuscript, limited discussion was devoted to this aspect within the text. (**Page 23, Lines 515-527**)

5. As for Figure 5A, it would be better to show enrichment results for all examined tissues.

RESPONSE: We appreciate the reviewer's suggestion. To evaluate the functional enrichment of puQTLs, we procured chromatin state data from various tissues as documented in the Roadmap Epigenomics project. It is important to note that certain tissue types from the GTEx study lack corresponding data in the Roadmap Epigenomics project. Our findings regarding enrichment are presented for all tissues with available chromatin state data in the Roadmap Epigenomics project.

Reviewer #2 (Remarks to the Author):

The authors have answered several of my queries, however there are still major points that require clarification.

RESPONSE: We extend our sincere appreciation to the reviewer for the insightful comments and valuable suggestions. The feedback provided by the reviewer has been carefully considered, leading to additional analyses being implemented in the revised manuscript. A detailed response addressing the reviewers' comments is presented.

Major points

Point 1: I thank the authors for their clarification of the filtering procedure and have no further comments.

RESPONSE: We thank for the valuable comment again.

Point 2: I remain very confused with the authors tally of the number of loci discovered and their methodology for calculating multiple puQTLs for a single major promoter. There are several places in the text where authors use the terms ‘puQTL loci’ and ‘genetic variant’ interchangeably, suggesting they are mis-using these terms, and counting every genetic variant with an association below their thresholds as an individual puQTL, rather than controlling for linked SNPs within a locus. This has serious implications for all subsequent analyses.

For example, the response states: ″Subsequent to the expansion of our dataset, the count of unique puQTL loci has been revised to 602,009″. But in the revised abstract it states: ″By constructing an atlas of human puQTLs across 4935 different tissues from 838 individuals, we have identified approximately 602,009 genetic variants associated with promoter usage.″

The authors should clearly state how many independent loci these 602,009 genetic variants correspond to.

RESPONSE: We extend our sincere appreciation to the reviewer for highlighting a pivotal concern. In this study, a total of 602,009 genetic variants have been identified as puSNPs across 49 tissues, corresponding to 76,856 LD-independent loci (**Supplementary Data 1**). The misapplication of the term "independent loci" has been rectified in the revised manuscript. In order to distinguish the “genetic variants” and “loci”, the nomenclature utilized in this study designates these variants as "puSNP", "paSNP" or “eSNP”, while the corresponding associations are denoted as "puSNP-gene pairs" or "puQTL pairs", “paSNP-gene pairs” or “paQTL pairs”, “eSNP-gene pairs” or “eQTL pairs” throughout the entirety of the manuscript. (**Page 7, Lines 131-134; Page 10, Lines 201-202**)

Point 3: Related to point 1 above, I am very worried that the authors are conflating discovery of multiple independent puQTLs with the presence of multiple SNPs in LD within a single loci/puQTL. This discrepancy must be resolved and corrected prior to any publication.

For example, the authors state in the response:
However, it's worth noting that while there's only one major promoter for a given gene, there can be multiple puQTLs independently associated with the usage of this major promoter. For example, in the case of gene DUSP18, the promoter 49007 serves as the major promoter, and the study identified five puQTLs significantly linked to the usage of this specific promoter (Figure S7). We

have drawn a model to elucidate the regulatory relationships between multiple puQTLs and the major promoter (Response Figure 5). Additionally, we have clarified the relationships between multiple puQTLs and the major promoter in the Results and Methods section. (Pages 18-19, Lines 407-420; Page 25, Lines 583-585)

The referenced text from line 407 states:
In a similar vein, we identified a group of five adjacent SNPs in conjunction with promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase (DUSP) family, which serve as critical modulators subsequent to acute lung injury. We observed long-range H3K27ac mediated chromatin interactions between the SNPs and promoter 49007, and the promoter was bound by H3K27ac411 and H3K4me1 in lung tissue (Figure S7A). The puQTLs rs5749148 and rs5753260 displayed significant associations with the upregulation of promoter 49007, while puQTLs rs5753259, rs5749147, and rs4820023 were significantly linked to the downregulation of its usage (Figures S7B and S7C). Furthermore, the aforementioned five SNPs, characterized as potential paQTLs of promoter 49007, did not exhibit a significant effect on the activity of the promoter itself.

However, the example ‘independent’ SNPs rs5749148 and rs5753260 are in near complete LD (R2 = 0.996 in EUR), and in very high LD with the remaining three SNPs (R2= 1.0, 0.8 and 0.69). I cannot see how these five tightly linked SNPS could possibly represent five “multiple puQTLs independently associated with the usage of this major promoter.” If a claim of multiple independent puQTLs for a single promotor are to be included in the manuscript the authors MUST perform analysis to demonstrate they are independent and not merely one signal that is being tagged by multiple linked SNPs. The authors should also clarify if subsequent analysis throughout the paper including all 602,009 genetic variants or if they have used a reduced set representing independent SNPs (ie one lead SNP per independent loci), and how many total independent puQTLs they identify.

RESPONSE: As we have responded in comment #2, the total of 602,009 genetic variants identified were associated with 76,856 LD-independent loci. These independent loci were utilized in comparative analyses with lead eQTLs (**Fig. 3c, d, e**). Subsequent analyses were carried out using the complete dataset comprising all 602,009 genetic variants. A comprehensive overview of the independent loci across different tissues has been documented in **Supplementary Data 1**. Furthermore, detailed information regarding the designation of a puQTL as a lead SNP has been

meticulously curated and made accessible through the data portal (http://bioailab.com:3838/GTEX_puQTL/) for reference and exploration.

Besides, it is important to note that the genotypes of SNPs have been reformatted in all figures, utilizing the reference and alternative allele in place of the traditional "AA", "Aa", and "aa" nomenclature throughout the manuscript. As a result, it is possible that the orientation of ß values may be altered accordingly. So that, in the revised results, the five puSNPs were associated with the upregulation of the promoter 49007 usage in *DUSP18* gene. (**Pages 18-19, Lines 417-422**)

Point 4: I thank the reviewer for their response to my original point. However, given their clarification, I think the following manuscript statement is misleading: However, as less than 16% of overlapped genes were associated with the same SNPs in both puQTLs and eQTLs, suggesting the presence of unique associations between these two types of QTLs and their respective targets (Figure3C).”Please expand the above results statement to either clearly state it only refers to the comparison on LEAD eQTL SNPs, and add a caveat explaining that the 16% overlap is likely a large underestimate as the lead SNP can vary considerably in regions with high LD.

RESPONSE: We acknowledge the insightful feedback provided by the reviewer. The manuscript has been revised to explicitly specify that the comparison was carried out for lead SNPs, and the potential factors contributing to the underestimated overlapping ratio have been thoroughly discussed. (**Page 12, Lines 249-254**)

Point 5 – I appreciate the additional analyses provided by the authors. I would suggest Response Figure 6 and the reported overall effect size correlation should be included in the manuscript supplemental material

RESPONSE: Thanks for the constructive suggestion. We have added the overall effect size results into the manuscript (**Supplementary Fig. 4**). (**Page 12, Lines 268-269**)

Point 6 – I thank the authors for their clarification

RESPONSE: We thank for the valuable comment again.

Point 7 - I thank the authors for their clarification

RESPONSE: We thank for the valuable comment again.

Point 8 – I thank the authors for their comprehensive response, but I believe they have misunderstood my query. I am uncertain if this is a simple mis-understanding of their description of their pipeline on my part or a methodological mistake, and I request the authors to clarify carefully.

The authors description of their coloc analyses has left me worried that they ran the coloc package on the subset of SNPs that met a significance threshold rather than including ALL SNPs in the genomic region that contains the significant loci. For example, the new methods text states, “To conduct the colocalization analysis of puQTL-promoter pairs, we first obtained the significant SNPs from each GWAS trait (P-value < 1 x 10-5) and subsequently compared them with the complete puQTL set, encompassing SNPs located within a 1Mb proximity of the interested promoter.” This description implies that only SNPs which had been either p<10-5 in a GWAS trait or called as a puQTL were included in the co-localization.

While I appreciate the Response Table 1, the content lists the significance cut off values others studies have used for determining which loci to colocalize. These studies subsequently used ALL SNPs in the genomic regions that they co-localized. I would like the authors to be clear that in the actual co-localization between a puQTL loci and a GWAS they used ALL SNPs in the genomic region – not just the SNPs that passed a threshold. Trimming the input SNPS by a significance threshold will invalidate coloc.

I have pasted the instructions from the coloc website below:
First: select the SNPs to study. SNPs must * have summary data available in both studies * cover a single genomic region (perhaps defined by distance about some GWAS peak or delimited by recombination hotspots), and * represent a dense coverage of the region

Please do include all SNPs you have in the region: do not trim by significance or MAF or other means. Coloc enumerates all possible configurations of shared and non-shared causal variants and evaluates their relative support; it cannot do this if it does not have all the SNPs available.

RESPONSE: We appreciated the constructive feedback provided by the reviewer which significantly enhanced our understanding of the colocalization analysis. We agreed with the reviewer that all SNPs should be included for the colocalization analysis. Subsequently, we have

revised the co-localization analysis by encompassing all SNPs within a 1Mb region of the target promoter region. We have extracted all SNPs within this region associated with each GWAS trait and subsequently compared them with the complete puQTL set. Following the revision of our findings, corresponding figures have been updated accordingly (**Fig. 6, Supplementary Fig. 9, 10**). (**Page 19, Lines 430-434; 439-440; Page 20, Lines 446-461; Page 32, Lines 741-744**)

Point 9: I thank the authors for their methods clarification

RESPONSE: We thank for the valuable comment again.

Point 10: I thank the authors for adding these data availability resources

RESPONSE: We thank for the valuable comment again.

Reviewer #3 (Remarks to the Author):

The authors have addressed all of my concerns about the content of the paper. However, I would still strongly recommend also publicly releasing the full promoter usage QTL summary statistics. This would allow other research to use the results from this paper in their analysis and would significantly increase the impact of the paper.

RESPONSE: We would like to express our gratitude for the valuable recommendation provided by the reviewer. Efforts have been made to address the matter of publicly sharing of the full summary statistics data. The full summary statistics data can now be accessed through the provided links on the data portal [http://bioailab.com:3838/GTEX puQTL/], as mentioned in the data availability section.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have properly addressed all my concerns. Their comparison between puQTL and sQTL was particularly clear.

In the provided link (http://bioailab.com:3838/GTEX_puQTL/), I was able to locate a pQTL full summary statistics file for just one tissue (Adipose-Subcutaneous.gz). I would highly suggest making summary statistics for all tissues available to the public.

Reviewer #2 (Remarks to the Author):

I have responded to some of the authors response below, and raised queries on one new statements added to this version of the manuscript. Overall, the manuscript contained several changes that were not tracked, so it was hard to review all changes, and I have not exhaustively checked for corrections to the the puSNP vs puQTL reporting. Points from the previous review that are not included below have been answered by the authors.

RESPONSE: We extend our sincere appreciation to the reviewer for highlighting a pivotal concern. In this study, a total of 602,009 genetic variants have been identified as puSNPs across 49 tissues, corresponding to 76,856 LD-independent loci (Supplementary Data 1). The misapplication of the term "independent loci" has been rectified in the revised manuscript. In order to distinguish the “genetic variants” and “loci”, the nomenclature utilized in this study designates these variants as "puSNP", "paSNP" or “eSNP”, while the corresponding associations are denoted as "puSNP-gene pairs" or "puQTL pairs", “paSNP-gene pairs” or “paQTL pairs”, “eSNP-gene pairs” or “eQTL pairs” throughout the entirety of the manuscript. (Page 7, Lines 131-134; Page 10, Lines 201-202)

I thank the authors for clarifying this point. This clarification however does raise many questions about the results and the original and revised manuscript.

First, I cannot find a description in the methods of how the 76,856 LD-independent loci were calculated. Please add this to the methods and state what LD threshold was used to identify an independent loci.

While ‘"puSNP-gene pairs" or similar is understandable, I do not understand what a ‘paQTL pairs’ is. Please define this term or remove it from the manuscript.

The analysis of and reporting of puSNPs throughout the manuscript instead of puQTLs or lead puSNPs is odd, and leads to confusion in several sections and difficulty in interpreting results, as it is unclear whether a result is influenced by the stronger LD in some regions in others (which will result in more linked puSNPs for a given puQTL) or the magnitude of the association (lower p value of the causal SNP will cause more puSNPs with decreasing LD to the causal variant to pass the eSNP thresholds). This will bias many of the downstream analyses

Please revise the abstract to clearly state the number of independent puQTLs discovered, rather than the number of linked genetic variants associated to a puQTL.

Within the text, when stating the number of puSNPs, please also be clear how many puQTLs this corresponds to. For example Line 134: The number of puSNPs varied across tissues, ranging from 999

in kidney cortex to 155,547 in thyroid tissues

Some tables still seem to be reporting the wrong number – for example supplementary Table 6 reports ‘puQTL-gene overlapped with paQTL-genes, with the first entry as 89,985. How can there be 89K genes?

Please clarify if the below statement refers to independent genetic variants: Notably, tissues with larger sample sizes, such as thyroid, sun-exposed skin and lung, displayed a high proportion of promoters with more than four variants.

Original Point 3: Related to point 1 above, I am very worried that the authors are conflating discovery of multiple independent puQTLs with the presence of multiple SNPs in LD within a single loci/puQTL. This discrepancy must be resolved and corrected prior to any publication.

For example, the authors state in the response:
However, it's worth noting that while there's only one major promoter for a given gene, there can be multiple puQTLs independently associated with the usage of this major promoter. For example, in the case of gene DUSP18, the promoter 49007 serves as the major promoter, and the study identified five puQTLs significantly linked to the usage of this specific promoter (Figure S7). We have drawn a model to elucidate the regulatory relationships between multiple puQTLs and the major promoter (Response Figure 5). Additionally, we have clarified the relationships between multiple puQTLs and the major promoter in the Results and Methods section. (Pages 18-19, Lines 407-420; Page 25, Lines 583-585)

The referenced text from line 407 states:
In a similar vein, we identified a group of five adjacent SNPs in conjunction with promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase (DUSP) family, which serve as critical modulators subsequent to acute lung injury. We observed long-range H3K27ac mediated chromatin interactions between the SNPs and promoter 49007, and the promoter was bound by H3K27ac411 and H3K4me1 in lung tissue (Figure S7A). The puQTLs rs5749148 and rs5753260 displayed significant associations with the upregulation of promoter 49007, while puQTLs rs5753259, rs5749147, and rs4820023 were significantly linked to the downregulation of its usage (Figures S7B and S7C). Furthermore, the aforementioned five SNPs, characterized as potential paQTLs of promoter 49007, did not exhibit a significant effect on the activity of the promoter itself.

However, the example ‘independent’ SNPs rs5749148 and rs5753260 are in near complete LD (R2 = 0.996 in EUR), and in very high LD with the remaining three SNPs (R2= 1.0, 0.8 and 0.69). I cannot see how these five tightly linked SNPS could possibly represent five “multiple puQTLs independently associated with the usage of this major promoter.” If a claim of multiple independent puQTLs for a single promotor are to be included in the manuscript the authors MUST perform analysis to demonstrate they are independent and not merely one signal that is being tagged by multiple linked SNPs. The authors should also clarify if subsequent analysis throughout the paper including all 602,009 genetic variants or if they have used a reduced set representing independent SNPs (ie one lead SNP per independent loci), and how many total independent puQTLs they identify.

RESPONSE: As we have responded in comment #2, the total of 602,009 genetic variants identified were associated with 76,856 LD-independent loci. These independent loci were utilized in comparative analyses with lead eQTLs (Fig. 3c, d, e). Subsequent analyses were carried out using the complete dataset comprising all 602,009 genetic variants. A comprehensive overview of the independent loci across different tissues has been documented in Supplementary Data 1. Furthermore, detailed

information regarding the designation of a puQTL as a lead SNP has been meticulously curated and made accessible through the data portal (http://bioailab.com:3838/GTEX_puQTL/) for reference and exploration.

REVIEWER RESPONSE - The text has been modified to swap out puQTL for puSNP, but I do not understand why it still discusses five puSNPs that are in near perfect LD? This will imply to a reader that there are five independent SNPs here. Please clarify in the text that the five are in high LD or edit to report the single puQTL at this locus.

Text: In a similar vein, we identified a group of five puSNPs in conjunction with promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase (DUSP) family, which serve as critical modulators subsequent to acute lung injury between the SNPs and promoter 49007, and the promoter was bound by H3K27ac and H3K4me1 in lung tissue (Supplementary Fig. 8a). The five puSNPs displayed significant associations with the upregulation of promoter 49007 usage, while did not
. We observed long-range H3K27ac mediated chromatin interactions exhibit a significant effect on the activity of the promoter (Supplementary Fig. 8b-f). However, they did demonstrate a notable impact on the activity of the nearby promoter 49006, despite the absence of H3K27ac-mediated chromatin interactions between the SNP and promoter 49006. Collectively, these findings indicate that puQTLs, rather than paQTLs, can be employed to identify new functional promoters that possess additional epigenetic characteristics.

Original Point 4: I thank the author for their response to my original point. However, given their clarification, I think the following manuscript statement is misleading: However, as less than 16% of overlapped genes were associated with the same SNPs in both puQTLs and eQTLs, suggesting the presence of unique associations between these two types of QTLs and their respective targets (Figure3C).”Please expand the above results statement to either clearly state it only refers to the comparison on LEAD eQTL SNPs, and add a caveat explaining that the 16% overlap is likely a large underestimate as the lead SNP can vary considerably in regions with high LD.

RESPONSE: We acknowledge the insightful feedback provided by the reviewer. The manuscript has been revised to explicitly specify that the comparison was carried out for lead SNPs, and the potential factors contributing to the underestimated overlapping ratio have been thoroughly discussed. (Page 12, Lines 249-254)

New text from manuscript "However, less than 16% of overlapped genes were associated with the same lead SNPs in both puQTLs and eQTLs, suggesting the presence of unique associations between these two types of QTLs and their respective targets (Fig. 3c). It is important to note that the overlapping ratios between lead puQTLs and eQTLs may have been underestimated due to substantial variation in lead SNP locations within regions exhibiting high linkage disequilibrium (LD)."

Reviewer response While I appreciate that it is now clear that you have only compared the lead SNPs, and added a brief caveat (not 'thoroughly discussed'), unless you have done some sort of LD filter to check if the lead SNPs are not in LD, I do not think this comparison is meaningful. I would suggest reporting the overlap to include all lead SNPs that in in LD of r2 of 0.6 or greater as a shared effect.

Additional text in this version: Notably, all these three types of QTLs displayed an excess of low P values in the GWAS of various traits, such as diastolic blood pressure in the artery aorta, multiple

sclerosis in the brain cortex, and high-density lipoprotein in the adipose subcutaneous

Reviewer response: It is well known that neither GWAS hits or eQTLs are randomly distributed across the genome. Both are more likely to occur nearby genes which can cause a artificial inflation of low p values in analyses like the one detailed above. How have you controlled for genomic position in the above analysis?

Reviewer #3 (Remarks to the Author):

The authors have adressed all of my intial concerns. However, I have now reviewed the issues raised by the other reviewers and agree that those should be adressed as well.

Point-by-point address of the comments raised by the reviewers for manuscript NCOMMS-23-45003B

Reviewer #1 (Remarks to the Author):

1. The authors have properly addressed all my concerns. Their comparison between puQTL and sQTL was particularly clear.

RESPONSE: We are grateful for the positive feedback provided by the reviewer, which has greatly enriched the study. The constructive comments offered by the reviewer have significantly improved the quality of the manuscript.

2. In the provided link (http://bioailab.com:3838/GTEX_puQTL/), I was able to locate a pQTL full summary statistics file for just one tissue (Adipose-Subcutaneous.gz). I would highly suggest making summary statistics for all tissues available to the public.

RESPONSE: Thanks for your constructive comments, On the data portal (http://bioailab.com:3838/GTEX_puQTL/), you can access the puQTL result for different tissue types by choosing from the dropdown menu and then clicking the submit button. The portal includes links to download full summary statistics for all tissues. Additionally, we have added a concise tip on searching and navigating the data portal.

Reviewer #2 (Remarks to the Author):

I have responded to some of the authors response below, and raised queries on one new statements added to this version of the manuscript. Overall, the manuscript contained several changes that were not tracked, so it was hard to review all changes, and I have not exhaustively checked for corrections to the the puSNP vs puQTL reporting. Points from the previous review that are not included below have been answered by the authors.

RESPONSE: We appreciate the reviewer for the evaluation. Your feedback is valuable in enhancing the quality and significance of the study. The updated text in the revised manuscript has been marked in red, with the relevant page and line numbers noted at the end of each response paragraph. Given that you have referenced some of our previous responses within you new comments, we have applied a gray shading to distinguish those earlier responses, while your current round of comments is in back, and our replies are in blue. To aid you in navigating the updated content with ease, we have embedded the revised text within each point-by-point response.

Comment 1:

RESPONSE: We extend our sincere appreciation to the reviewer for highlighting a pivotal concern. In this study, a total of 602,009 genetic variants have been identified as puSNPs across 49 tissues, corresponding to 76,856 LD-independent loci (Supplementary Data 1). The misapplication of the term "independent loci" has been rectified in the revised manuscript. In order to distinguish the “genetic variants” and “loci”, the nomenclature utilized in this study designates these variants as "puSNP", "paSNP" or “eSNP”, while the corresponding associations are denoted as "puSNP-gene pairs" or "puQTL pairs", “paSNP-gene pairs” or “paQTL pairs”, “eSNP-gene pairs” or “eQTL pairs” throughout the entirety of the manuscript. (Page 7, Lines 131-134; Page 10, Lines 201-202).

I thank the authors for clarifying this point. This clarification however does raise many questions about the results and the original and revised manuscript. First, I cannot find a description in the methods of how the 76,856 LD-independent loci were calculated. Please add this to the methods and state what LD threshold was used to identify an independent loci.

RESPONSE: We appreciate the reviewer's reminder. The description of LD-independent loci filtering has been added to the methods section, and the parameter is determined based on other

relevant studies (Lim Y, et al., *PNAS.,* 2018, PMID: 30463956). The method corresponding to this is revised as follows:

"*We employed the 'ld_clump' function within the R package plinkbinr to identify LD-independent loci. By utilizing PLINK's '--clump' feature for LD clumping, we applied an LD threshold of r^2=0.1 within a 2 Mb window to exclude SNPs in strong LD (command: --clump-r^2 0.1 --clump-kb 2000). The LD calculations were conducted based on the reference panel of the 1000 Genomes Project.*" **(Page 22, Lines 574-579)**

Comment 2:

While '"puSNP-gene pairs" or similar is understandable, I do not understand what a 'paQTL pairs' is. Please define this term or remove it from the manuscript.

RESPONSE: We appreciate the reviewer for bringing up this concern. To address potential confusion regarding terminology, we have replaced the term "paQTL pairs" with "paSNP-gene pairs". This revision aims to enhance clarity and ensure consistency in terminology across the document.

Comment 3:

The analysis of and reporting of puSNPs throughout the manuscript instead of puQTLs or lead puSNPs is odd, and leads to confusion in several sections and difficulty in interpreting results, as it is unclear whether a result is influenced by the stronger LD in some regions in others (which will result in more linked puSNPs for a given puQTL) or the magnitude of the association (lower p value of the causal SNP will cause more puSNPs with decreasing LD to the causal variant to pass the eSNP thresholds). This will bias many of the downstream analyses.

RESPONSE: Thanks for your comments. We recognize that a QTL represents a region of the genome containing genetic variants that influence a quantitative trait, particularly promoter usage in our study. A QTL can encompass multiple SNPs, including the lead SNP and other variants contributing to the trait's heritability. To avoid any confusion in terminology, we have revised our manuscript to use "puSNP" to refer to specific variants.

Our study incorporates both lead puSNPs and LD puSNPs in several analyses to identify functional variants. The lead SNP, typically identified based on statistical significance, serves as an index for the QTL and may not be the causal SNP itself, as lead SNPs can be in LD with the actual causal variants. Given the biological significance and potential functionality of LD SNPs,

our analyses include both lead puSNPs and LD puSNPs in several analyses to pinpoint functional variants. This approach is consistent with current scientific practices, as evidenced in QTL studies on (Liang D, et al., *Nat Neurosci.,* 2018, PMID: 34017130; Garieri M, et al., *Nat Commun.,* 2017, PMID: 29116076; Kubota N, et al., *PLoS Comput Biol.,* 2022, PMID: 36037215), all of which have utilized both the lead SNPs and LD SNPs for their analyses.

We noted the reviewer's concern regarding potential bias from including LD puSNPs. We conducted a thorough examination of all analyses to ensure unbiased findings. Additional analyses from lead puSNPs corroborate our results, including statistics and distributions of puQTLs (**Fig. 1 and Supplementary Fig. 1**), the comparison of puQTLs with eQTLs (**Fig. 3**) and the shared tissue analysis (**Fig. 4 and Supplementary Fig. 7**). These parts are descriptive, and we have taken care to clarify the results to mitigate any misunderstanding. **(Page 6, Lines 135-137; Pages 11-12, Lines 289-291)**

Moreover, we have reviewed several downstream analyses, including tissue specificity analysis using mash, epigenomic features enrichment analysis using GREGOR, colocalization analysis, and partitioned heritability analysis. For the tissue specificity analysis with mash, we followed the guidelines. The development of mash accounted for LD (Urbut et al. *Nat Genet.,* 2019, PMID: 30478440). For the epigenomic features enrichment analysis by GREGOR, we adhere to the guidelines. GREGOR was designed to quantify the enrichment of both lead SNPs and LD SNPs, considering that the lead SNP is not necessarily be the causal variant due to LD (Schmidt E et al., *Bioinformatics,* 2015, PMID: 25886982). For colocalization analysis, we included all SNPs as per your previous comments. In the partitioned heritability analysis, we included both lead puSNPs and LD puSNPs to accurately estimate heritability using LDSC. LDSC is designed to handle the entire spectrum of SNP effects and leverages the comprehensive LD structure for robust heritability estimation (Finucane H et al., *Nat Genet.,* 2015, PMID: 26414678). Including both lead puSNPs and LD puSNPs captures the full spectrum of genetic regulation, providing a more complete picture of heritability. Limiting the analysis to lead puSNPs might miss additional regulatory variants that contribute the molecular trait and, subsequently, to trait heritability.

Collectively, we believe our findings are not biased due to LD and provide a comprehensive understanding of the genetic regulation involved in alternative promoters.

Comment 4:

Please revise the abstract to clearly state the number of independent puQTLs discovered, rather than the number of linked genetic variants associated to a puQTL.

RESPONSE: We appreciate the reviewer's feedback. As we adjustedWe have revised the abstract to clearly state the number of independent loci discovered as follows:

"*By constructing an atlas of human puQTLs across 49 different tissues from 838 individuals, we have identified approximately 76,856 independent loci associated with promoter usage, encompassing 602,009 genetic variants.*" **(Page 2, Lines 33-36)**

Comment 5:

Within the text, when stating the number of puSNPs, please also be clear how many puQTLs this corresponds to. For example, Line 134: The number of puSNPs varied across tissues, ranging from 999 in kidney cortex to 155,547 in thyroid tissues.

RESPONSE: We appreciate the comment. We have revised the text accordingly as follows:

"*The number of genetic variants associated with promoter usage varied across tissues, ranging from 999 in kidney cortex to 155,547 in thyroid tissues, corresponding to 42 and 5,328 independent loci in kidney and thyroid respectively.*" **(Pages 5-6, Lines 127-129)**

Comment 6:

Some tables still seem to be reporting the wrong number – for example supplementary Table 6 reports 'puQTL-gene overlapped with paQTL-genes, with the first entry as 89,985. How can there be 89K genes?

RESPONSE: We apologize for the confusion. The values reported in Supplementary Table 6 refer to the puSNP-gene pairs and paSNP-gene pairs. Thus, it is possible for the number of pairs to be greater than the number of unique genes. We have updated the column in Supplementary Table 6 to clearly reflect this distinction.

Comment 7:

Please clarify if the below statement refers to independent genetic variants: Notably, tissues with larger sample sizes, such as thyroid, sun-exposed skin and lung, displayed a high proportion of promoters with more than four variants.

RESPONSE: The genetic variants mentioned here are not independent loci, but rather puSNPs. In response to this feedback, we have revised the analysis of the independent loci to demonstrate

alignment with the results observed at the puSNP level (**Supplementary Fig. 1e**). To avoid any confusion, we have revised the text accordingly as follows:

"*Notably, tissues with larger sample sizes, such as thyroid, sun-exposed skin and lung, displayed a high proportion of promoters with multiple independent loci (**Supplementary Fig. 1e**)*". **(Page 6, Lines 135-137)**

Comment 8:

Original Point 3: Related to point 1 above, I am very worried that the authors are conflating discovery of multiple independent puQTLs with the presence of multiple SNPs in LD within a single loci/puQTL. This discrepancy must be resolved and corrected prior to any publication.

For example, the authors state in the response:

However, it's worth noting that while there's only one major promoter for a given gene, there can be multiple puQTLs independently associated with the usage of this major promoter. For example, in the case of gene DUSP18, the promoter 49007 serves as the major promoter, and the study identified five puQTLs significantly linked to the usage of this specific promoter (Figure S7). We have drawn a model to elucidate the regulatory relationships between multiple puQTLs and the major promoter (Response Figure 5). Additionally, we have clarified the relationships between multiple puQTLs and the major promoter in the Results and Methods section. (Pages 18-19, Lines 407-420; Page 25, Lines 583-585)

The referenced text from line 407 states: In a similar vein, we identified a group of five adjacent SNPs in conjunction with promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase (DUSP) family, which serve as critical modulators subsequent to acute lung injury. We observed long-range H3K27ac mediated chromatin interactions between the SNPs and promoter 49007, and the promoter was bound by H3K27ac411 and H3K4me1 in lung tissue (Figure S7A). The puQTLs rs5749148 and rs5753260 displayed significant associations with the upregulation of promoter 49007, while puQTLs rs5753259, rs5749147, and rs4820023 were significantly linked to the downregulation of its usage (Figures S7B and S7C). Furthermore, the aforementioned five SNPs, characterized as potential paQTLs of promoter 49007, did not exhibit a significant effect on the activity of the promoter itself.

However, the example ‘independent’ SNPs rs5749148 and rs5753260 are in near complete LD (R2 = 0.996 in EUR), and in very high LD with the remaining three SNPs (R2= 1.0, 0.8 and 0.69). I cannot see how these five tightly linked SNPS could possibly represent five “multiple puQTLs independently associated with the usage of this major promoter.” If a claim of multiple independent puQTLs for a single promotor are to be included in the manuscript the authors MUST perform analysis to demonstrate they are independent and not merely one signal that is being tagged by multiple linked SNPs. The authors should also clarify if subsequent analysis throughout the paper including all 602,009 genetic variants or if they have used a reduced set representing independent SNPs (ie one lead SNP per independent loci), and how many total independent puQTLs they identify.

RESPONSE: As we have responded in comment #2, the total of 602,009 genetic variants identified were associated with 76,856 LD-independent loci. These independent loci were utilized in comparative analyses with lead eQTLs (Fig. 3c, d, e). Subsequent analyses were carried out using the complete dataset comprising all 602,009 genetic variants. A comprehensive overview of the independent loci across different tissues has been documented in Supplementary Data 1. Furthermore, detailed information regarding the designation of a puQTL as a lead SNP has been meticulously curated and made accessible through the data portal (http://bioailab.com:3838/GTEX_puQTL/) for reference and exploration.

Text: In a similar vein, we identified a group of five puSNPs in conjunction with promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase (DUSP) family, which serve as critical modulators subsequent to acute lung injury between the SNPs and promoter 49007, and the promoter was bound by H3K27ac and H3K4me1 in lung tissue (Supplementary Fig. 8a). The five puSNPs displayed significant associations with the upregulation of promoter 49007 usage, while did not. We observed long-range H3K27ac mediated chromatin interactions exhibit a significant effect on the activity of the promoter (Supplementary Fig. 8b-f). However, they did demonstrate a notable impact on the activity of the nearby promoter 49006, despite the absence of H3K27ac-mediated chromatin interactions between the SNP and promoter 49006. Collectively, these findings indicate that puQTLs, rather than paQTLs, can be employed to identify new functional promoters that possess additional epigenetic characteristics.

REVIEWER RESPONSE - The text has been modified to swap out puQTL for puSNP, but I do not understand why it still discusses five puSNPs that are in near perfect LD? This will imply to a

reader that there are five independent SNPs here. Please clarify in the text that the five are in high LD or edit to report the single puQTL at this locus.

RESPONSE: Thank you for the comments. Fine-mapping causal variants in a genetic locus is essential for understanding the underlying mechanisms by which a variant can contribute to a trait. Overcoming LD and pinpointing context-specific causal variants is imperative. Many functional and computational fine-mapping methods have been developed for this purpose. One straightforward approach is to overlap variants with functional elements. In our study, we employed different histone data to determine the functional elements.

We found that the five puSNPs were bound by the enhancer-associated histone marks H3K27ac and H3K4me1, suggesting their potential location within enhancer regions. Additionally, long-range chromatin interactions were observed between this region and the promoter 49007 of *DUSP18*. Therefore, discussing this region is warranted. Given that a lead SNP shows the strongest statistical associations and is not necessarily the causal variant, and that previous studies have shown a substantial proportion of SNPs in high LD with the lead SNP exhibit higher stronger causal evidence that the lead SNP itself (Schaub MA, et al. *Genome Res.* 2012. PMID: 22955986), we initially incorporated all five LD-linked SNPs. Given the linkage disequilibrium among these SNPs, discerning the causal SNPs based on effect size or association significance of SNP genotypes was challenging. However, fine-mapping analysis using SuSiE identified rs5753260 as the causal variant.

To avoid misunderstanding, we have clarified that these five SNPs are in high LD. Specifically, we revised the boxplot to display only SNP rs5753260 in Supplementary Figure 10, alongside the fine-mapping results. We have revised the manuscript as follows:

"*Furthermore, we identified a cluster of five SNPs in high LD associated with the promoter 49007 of the DUSP18 gene, an atypical member of the dual-specificity phosphatase family that acts as a critical modulator subsequent to acute lung injury. These puSNPs were marked by H3K27ac and H3K4me1, suggesting their presence within enhancer regions. Additionally, long-range H3K27ac-mediated chromatin interactions were observed between these SNPs and the promoter 49007 in lung tissue (**Supplementary Fig. 10a**). These puSNPs exhibited significant associations with the upregulation of promoter 49007 usage. SuSiE fine-mapping analysis identified rs5753260 as the causal variant (**Supplementary Fig. 10b, Supplementary Fig. 10c**). Collectively, rs5753260 may influence enhancer activity, therefore affecting the promoter usage of DUSP18 through enhancer-promoter interaction. This underscores the potential of integrating*

multi-omics data to identify functionally relevant genetic variants marked by epigenetic marks and long-range chromatin interactions." **(Page 15, Lines 384-395)**

Comment 9:

Original Point 4: I thank the author for their response to my original point. However, given their clarification, I think the following manuscript statement is misleading: However, as less than 16% of overlapped genes were associated with the same SNPs in both puQTLs and eQTLs, suggesting the presence of unique associations between these two types of QTLs and their respective targets (Figure3C)."Please expand the above results statement to either clearly state it only refers to the comparison on LEAD eQTL SNPs, and add a caveat explaining that the 16% overlap is likely a large underestimate as the lead SNP can vary considerably in regions with high LD.

RESPONSE: We acknowledge the insightful feedback provided by the reviewer. The manuscript has been revised to explicitly specify that the comparison was carried out for lead SNPs, and the potential factors contributing to the underestimated overlapping ratio have been thoroughly discussed. (Page 12, Lines 249-254)

New text from manuscript "However, less than 16% of overlapped genes were associated with the same lead SNPs in both puQTLs and eQTLs, suggesting the presence of unique associations between these two types of QTLs and their respective targets (Fig. 3c). It is important to note that the overlapping ratios between lead puQTLs and eQTLs may have been underestimated due to substantial variation in lead SNP locations within regions exhibiting high linkage disequilibrium (LD)."

Reviewer response: While I appreciate that it is now clear that you have only compared the lead SNPs and added a brief caveat (not 'thoroughly discussed'), unless you have done some sort of LD filter to check if the lead SNPs are not in LD, I do not think this comparison is meaningful. I would suggest reporting the overlap to include all lead SNPs that in in LD of r2 of 0.6 or greater as a shared effect.

RESPONSE: We appreciate the reviewer's feedback. In response to the concern regarding the meaningfulness of comparing only the lead SNPs without considering their linkage disequilibrium (LD), we have performed an additional analysis. Specifically, we assessed the percentage of

shared SNPs linked to the same gene between puQTLs and eQTLs, incorporating puQTLs that either share the same SNP ID as eQTLs or are in high LD ($r^2 \geq 0.6$) with eQTLs. This approach ensures that we account for genetic variants that may influence the same gene through LD relationships. Our results indicate that approximately 16% of shared genes are associated with identical SNPs in both puQTLs and eQTLs (Fig. 3c). By including puQTLs in high LD with eQTLs, we observed an increase in the proportion of shared target genes to 22%. We have revised the manuscript to reflect these updated findings as follows:

"*Approximately 16% of shared genes were found to be linked to identical SNPs in both puQTLs and eQTLs (**Fig. 3c**). Some puQTLs may be located in regions with high LD to eQTLs. Additionally, when puQTLs in high LD with eQTLs ($r^2 \geq 0.6$) are considered, the percentage of shared target genes increases to 22%. This indicates distinctive relationships between these QTL types and their corresponding targets (**Supplementary Fig. 5a**).*" **(Pages 9-10, Lines 234-239)**

Comment 10:

Additional text in this version: Notably, all these three types of QTLs displayed an excess of low P values in the GWAS of various traits, such as diastolic blood pressure in the artery aorta, multiple sclerosis in the brain cortex, and high-density lipoprotein in the adipose subcutaneous.

Reviewer response: It is well known that neither GWAS hits or eQTLs are randomly distributed across the genome. Both are more likely to occur nearby genes which can cause a artificial inflation of low p values in analyses like the one detailed above. How have you controlled for genomic position in the above analysis?

RESPONSE: We appreciate the reviewer's comment regarding the non-random distribution of GWAS hits and eQTLs across the genome. To address this concern, we would like to clarify our methodology. In our study, we employed quantile-quantile (Q-Q) plots to evaluate the enrichment of GWAS signals within different QTL types, including puQTLs, paQTL, and eQTL. Specially, GWAS SNPs overlapping with puQTLs were identified, and the Q-Q plot of GWAS *P*-values for these SNPs was generated. A similar approach was applied to GWAS SNPs overlapping with paQTLs and eQTLs. As a control, we also plotted the Q-Q plots of genome-wide GWAS *P*-values (**represented as black dots in Figure 6A and 6B**). By including genome-wide GWAS Q-Q plots as a baseline, we provided a reference to distinguish the enrichment signal specific to QTL overlaps from the general distribution of genome-wide GWAS P-values. (**Page 26, Lines 686-691**)

This approach is consistent with methodologies widely employed in QTL studies. For example, similar approaches were used in studies by the GTEx Consortium (*Science*, 2015, PMID: 25954001), Liu S, et al. (*Nat Genet.,* 2022, PMID: 35953587), Zhang Z, et al. (*Nat Genet.,* 2020, PMID: 32601472), Chen H, et al. (*Nat Commun.,* 2024, PMID: 38409266), Zhang M, et al. (*Gastroenterology,* 2023, PMID: 37541527), Hamel A, et al. (*Nat Commun.,* 2024, PMID: 38195602). Specifically, in the m6A QTL study (Zhang Z, et al., *Nat Genet.,* 2020, PMID: 32601472), GWAS SNPs overlapping with m6A QTLs was selectively analyzed using Q-Q plots, similar to our approach.

Reviewer #3 (Remarks to the Author):

The authors have addressed all of my initial concerns. However, I have now reviewed the issues raised by the other reviewers and agree that those should be adressed as well.

RESPONSE: We thank for the valuable comment again.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

The authors have now addressed my major concerns.

Point-by-point address of the comments raised by the reviewers for manuscript NCOMMS-23-45003C

We sincerely thank the esteemed editor and dedicated reviewers for their thorough evaluation of our revised manuscript. Their valuable insights have significantly improved the quality of our work. Below is our detailed point-by-point response to the reviewers' comments.

Reviewer #2 (Remarks to the Author):

The authors have now addressed my major concerns.

RESPONSE: We sincerely appreciate your time and efforts in reviewing our manuscript, as well as your insightful comments.