

Genome Analysis

PETScan: score-based genome-wide association analysis of RNA-Seq and ATAC-Seq data

Yajing Hao¹ , Tal Kafri^{2,3,4}, Fei Zou^{1,5,*} 

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

²Gene Therapy Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

³Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

⁴Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

⁵Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

*Corresponding author. Fei Zou, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: feizou@email.unc.edu

Associate Editor: Christina Kendzierski

Abstract

Motivation: High-dimensional sequencing data, such as RNA-Seq for gene expression and ATAC-Seq for chromatin accessibility, are widely used in studying systems biology. Accessible chromatin allows transcription factors and regulatory elements to bind to DNA, thereby regulating transcription through the activation or repression of target genes. The association analysis of RNA-Seq and ATAC-Seq data provides insights into gene regulatory mechanisms. Most existing analytic tools exclusively focus on cis-associations, despite regulatory elements being able to physically interact with distant target genes. Furthermore, conventional approaches often utilize Pearson or Spearman correlations, which ignore the count-based nature of RNA-Seq data.

Results: To address these limitations, we introduce *PETScan*, a computationally efficient genome-wide *PEak-Transcript Score*-based association analysis, utilizing negative binomial models to better accommodate RNA-Seq data. We leverage score tests and matrix calculations for improved computational efficiency, and combine an empirical permutation method with genomic control to ensure valid p-value calculations in studies with limited sample sizes. In real-world datasets, *PETScan* achieved three orders of magnitude faster than Wald tests, while identifying similar significant gene-peak pairs.

Availability: The *PETScan* R package is available on GitHub at <https://github.com/yajing-hao/PETScan>.

Introduction

Transcriptomics data analysis investigates RNA transcripts produced by the genome under specific biological conditions or in particular cell types, which plays a significant role in understanding complex traits and diseases (Casamassimi *et al.* 2017). Although every cell in the human body shares an identical set of genes, distinct gene expression patterns define the specialized functions of different cell types both physiologically and pathologically (Ralston and Shaw 2008, Zeng 2022). These patterns can be regulated by epigenetic mechanisms, such as DNA methylation (Dhar *et al.* 2021) and histone modifications (Gagnidze and Pfaff 2022), which regulate gene expression through modifications of DNA molecules or chromatin without altering the underlying DNA sequence. Additionally, changes to chromatin accessibility represent another important epigenetic mechanism that regulates gene expression (Tsompana and Buck 2014). Structurally, a chromatin region can be densely packed (heterochromatin), leading to gene silencing, while loosely packed chromatin (euchromatin) tends to be transcriptionally active (Huisinga *et al.* 2006). The dynamic nature of chromatin accessibility plays a crucial role in gene regulation by allowing transcription factors (TFs) and other regulatory elements to

bind to DNA, thereby regulating transcription through the activation or repression of target genes (Klemm *et al.* 2019).

High-throughput sequencing technologies, such as RNA sequencing (RNA-Seq) (Nagalakshmi *et al.* 2008) and Assay for Transposase-Accessible Chromatin sequencing (ATAC-Seq) (Buenrostro *et al.* 2013), enable the study of gene expression and chromatin accessibility, respectively. When RNA-Seq and ATAC-Seq data are jointly collected from the same samples or cells, systematic investigations of the relationship between gene expression and chromatin accessibility can be conducted to identify chromatin regions that regulate gene expression and reveal specifically impacted genes. Such analysis sheds light on the regulatory mechanisms underlying chromatin accessibility (Starks *et al.* 2019). Understanding these gene regulatory mechanisms is crucial for interpreting transcriptomic differences across various biological conditions, unraveling the progression of diseases (Kundaje *et al.* 2015), and developing effective therapeutic strategies (Dai *et al.* 2024). For example, research has identified disease-specific genes and uncovered their regulatory mechanisms underlying chromatin accessibility in diabetes (Ackermann *et al.* 2016), hepatocellular carcinoma (Yang *et al.* 2021, Bai *et al.* 2024), cellular senescence (Song *et al.* 2022, Ding *et al.* 2023), and asthma (Zhang *et al.* 2023). These studies provide

Received: 16 April 2025; Revised: 16 October 2025; Accepted: 16 November 2025

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

insights into how genetic and epigenetic factors contribute to disease development and progression.

To identify associations between gene expression and chromatin accessibility, most existing analyses rely on Pearson correlation (Zhu et al. 2019, Chen et al. 2019, Stuart et al. 2021) or more robust Spearman correlation (Ma et al. 2020, Kartha et al. 2022, Park and Rhee 2024). For instances, cisDyNet calculates Pearson correlation between gene expression and peak accessibility (Zhu et al. 2023), while TRIPOD detects peak-TF-gene trio regulatory relationships using Spearman correlation (Jiang et al. 2022). More advanced statistical methods include partial least-squares regression (Staitieh et al. 2023), LASSO (least absolute shrinkage and selection operator) (Cao et al. 2018), adaptive elastic-net regression (Ledru et al. 2024), and a tree-based non-linear regression method called gradient-boosting machine (Bravo González-Blas et al. 2023). All these methods ignore the count-based nature of RNA-Seq data, which can result in suboptimal performance and biased results, necessitating specialized statistical methods that appropriately handle RNA-Seq data.

Furthermore, current association analysis of RNA-Seq and ATAC-Seq data primarily focuses on identifying cis-associations in cis-regulatory regions (Frankel 2012, Klemm et al. 2019). For instance, cisDyNet evaluates the association between a gene's transcriptomic expression and peak accessibility within 500 kb upstream and downstream of its transcription start site (TSS) (Zhu et al. 2023). However, research has shown that gene regulatory elements bound to accessible chromatin can physically interact with distant target genes through chromatin looping (Kadauke and Blobel 2009), thereby modulating genes located far along the same chromosome or even on different chromosomes (Miele and Dekker 2008, Dean 2011, Dekker and Misteli 2015, Van Den Heuvel et al. 2015, Maass et al. 2019). A recent study has generated a comprehensive meta-Hi-C chromatin contact map, providing both functional cis- and trans-chromosomal interactions (Lohia et al. 2022), where trans-interactions refer to pairs of chromatin regions that are either located on different chromosomes or positioned far apart on the same chromosome.

To fully understand how chromatin accessibility regulates gene expression, a holistic approach that considers both cis- and trans-associations, or genome-wide association analysis, is necessary. However, RNA-Seq data typically includes expression profiles for over 20,000 genes, while ATAC-Seq data contains peak measurements from approximately 100,000 chromatin regions. This results in roughly $20,000 \times 100,000$, or 2 billion, paired association analyses, requiring a computationally efficient method that is currently unavailable.

To address these issues, we introduce *PETScan*, a computationally efficient genome-wide **PE**ak-**T**ranscript **S**core-based association analysis. The algorithm *PETScan* employs negative binomial models to respect the count nature of RNA-Seq data with potential overdispersion (Robinson et al. 2010, Di et al. 2011, Love et al. 2014). To mitigate the computational challenge, *PETScan* adopts two strategies to ensure its practical usage. First, *PETScan* utilizes score tests that require only parameter estimates under the null hypothesis. This feature becomes particularly useful since, for a given gene, its restricted estimates under the null hypothesis remain the same across all tested ATAC-Seq peaks and need to be evaluated only once.

Second, to further enhance computational efficiency, *PETScan* mimics Matrix eQTL (Shabalin 2012), an ultra-fast expression quantitative trait loci (eQTL) analysis tool. eQTL analysis also involves billions of paired association tests between millions of single-nucleotide polymorphisms (SNPs) and tens of thousands of genes across the entire genome. While Matrix eQTL converts traditional linear models into correlation coefficients for efficient large matrix operations, we combine score tests with ATAC-Seq data transformation to facilitate efficient matrix calculations.

Furthermore, in studies with small sample sizes, score tests may fail to converge to their theoretical asymptotic distributions. To better control the family-wise error rate or false discovery rate, *PETScan* incorporates an empirical permutation method inspired by genomic control (Devlin and Roeder 1999), which has been successfully used for robust statistical inference in differential gene expression analysis with RNA-Seq data (Zou et al. 2014). We demonstrate the effectiveness of *PETScan* using ENCODE human tissue-specific data and 10x Genomics embryonic mouse brain data. Computationally, *PETScan* is three orders of magnitude faster than the equivalent Wald tests.

Methods

Assume there are N paired samples of RNA-Seq and ATAC-Seq data with G genes and P peaks, respectively. Let $x_{1i}, x_{2i}, \dots, x_{(k-1)i}$ represent a set of covariates in the i th sample ($i = 1, \dots, N$), such as the top principal components derived from the RNA-Seq data, which are commonly used to account for batch effects. For gene g ($g = 1, \dots, G$) and peak p ($p = 1, \dots, P$), let y_{gi} be the expression level of gene g , and z_{pi} denote the log-transformed chromatin accessibility for peak p in the i th sample. For simplicity of notation, we drop g and p out of the subscripts when describing the model for testing the association between the gene g and peak p pair in the following discussion, whenever there is no confusion.

To adjust for potential over-dispersion, we assume that y_i follows a negative binomial distribution with the density function $f(y_i; \mu_i, \phi)$, where μ_i represents the mean and ϕ denotes the dispersion parameter with $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\phi}$. Specifically,

$$f(y_i; \mu_i, \phi) = \frac{\Gamma(y_i + \phi)}{y_i! \Gamma(\phi)} \left(\frac{\phi}{\phi + \mu_i} \right)^\phi \left(\frac{\mu_i}{\phi + \mu_i} \right)^{y_i}, \text{ where}$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_{(k-1)i} x_{(k-1)i} + \beta_k z_i = X_i \beta^* + \beta_k z_i,$$

$$\text{with } X_i = (1, x_{1i}, \dots, x_{(k-1)i}), \quad X = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix}, \quad \mathbf{z} = (z_1, \dots, z_N)^T,$$

$$\beta^* = (\beta_0, \beta_1, \dots, \beta_{(k-1)})^T, \beta = (\beta^{*T}, \beta_k)^T, \text{ and } \theta = (\phi, \beta^T)^T.$$

The log likelihood function of the data is therefore

$$l(\theta) = \sum_{i=1}^N \phi \log \left(\frac{\phi}{\phi + \mu_i} \right) + y_i \log \left(\frac{\mu_i}{\phi + \mu_i} \right) + \log[\Gamma(y_i + \phi)] - \log[\Gamma(\phi)]$$

with the corresponding score function $U(\theta) = \frac{\partial l(\theta)}{\partial \theta}$ and the Fisher information matrix $I(\theta) = E(-\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T})$. For the gene

and peak pair, the primary interest is to test $H_0 : \beta_k = 0$ versus $H_A : \beta_k \neq 0$, which assesses whether the chromatin accessibility at the peak affects the expression of the gene. Let $\tilde{\theta} = (\phi, \tilde{\beta}^{*T}, 0)^T$ be the restricted maximum likelihood estimate of θ under the null hypothesis. Note that the parameter estimate $\tilde{\theta}$ does not depend on the z_{pi} . Therefore, for gene g , it remains consistent across all P peaks and only needs to be evaluated once.

For testing $H_0 : \beta_k = 0$, *PETScan* uses the score test $SC = [U(\tilde{\theta})]_{\beta_k}^2 [I^{-1}(\tilde{\theta})]_{\beta_k \beta_k}$ which follows χ_1^2 under H_0 , where $U(\theta)_{\beta_k}$ denotes the $(k+2)$ th element of $U(\theta)$, and $[I^{-1}(\theta)]_{\beta_k \beta_k}$ represents the $(k+2, k+2)$ th entry of the inverse of $I(\theta)$, both corresponding to the parameter β_k . Specifically,

$$U(\tilde{\theta})_{\beta_k} = d_{\beta} \mathbf{z} \text{ and } I(\tilde{\theta}) = \begin{bmatrix} -\frac{\partial^2 l}{\partial \phi \partial \phi} \Big|_{\theta=\tilde{\theta}} & 0 & 0 \\ 0 & \mathbf{X}^T D_{\beta\beta} \mathbf{X} & \mathbf{X}^T D_{\beta\beta} \mathbf{z} \\ 0 & \mathbf{z}^T D_{\beta\beta} \mathbf{X} & \mathbf{z}^T D_{\beta\beta} \mathbf{z} \end{bmatrix},$$

where $d_{\beta} = \left(\frac{\tilde{\phi}(\bar{y}_1 - \bar{\mu}_1)}{\bar{\phi} + \bar{\mu}_1}, \dots, \frac{\tilde{\phi}(\bar{y}_N - \bar{\mu}_N)}{\bar{\phi} + \bar{\mu}_N} \right)$, $d_{\beta\beta} = \left(\frac{\tilde{\phi} \bar{\mu}_1 (\bar{\phi} + \bar{\mu}_1)}{(\bar{\phi} + \bar{\mu}_1)^2}, \dots, \frac{\tilde{\phi} \bar{\mu}_N (\bar{\phi} + \bar{\mu}_N)}{(\bar{\phi} + \bar{\mu}_N)^2} \right)$, and $D_{\beta\beta} = \text{diag}(d_{\beta\beta})$. Leveraging blockwise matrix inversion,

$$\begin{aligned} [I^{-1}(\tilde{\theta})]_{\beta_k \beta_k} &= (\mathbf{z}^T D_{\beta\beta} \mathbf{z} - \mathbf{z}^T D_{\beta\beta} \mathbf{X} (\mathbf{X}^T D_{\beta\beta} \mathbf{X})^{-1} \mathbf{X}^T D_{\beta\beta} \mathbf{z})^{-1} \\ &= (\mathbf{z}^{*T} D_{\beta\beta} \mathbf{z}^*)^{-1}, \end{aligned}$$

where $\mathbf{z}^* = [\mathbf{I} - \mathbf{X}(\mathbf{X}^T D_{\beta\beta} \mathbf{X})^{-1} \mathbf{X}^T D_{\beta\beta}] \mathbf{z}$. After constructing $\mathbf{z}^*$ by applying a transformation to $\mathbf{z}$, we can extract specific elements without computing the full inverse of a matrix.

Again, for gene g , d_{β} and $d_{\beta\beta}$ depend only on $\tilde{\theta}$ and remain consistent across all P peaks. Consequently, the score test for peak p can be extended easily to all peaks utilizing matrix calculations as described below: let $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_P]$, $\mathbf{Z}^* = [\mathbf{I} - \mathbf{X}(\mathbf{X}^T D_{\beta\beta} \mathbf{X})^{-1} \mathbf{X}^T D_{\beta\beta}] \mathbf{Z}$, then $\mathbf{SC} = (\mathbf{SC}_1, \dots, \mathbf{SC}_P) = (d_{\beta} \mathbf{Z}) \odot [d_{\beta\beta} (\mathbf{Z} \odot \mathbf{Z}^*)]^{\mathfrak{R}} \odot (d_{\beta} \mathbf{Z})$, where $\odot$ represents the Hadamard (element-wise) product and $A^{\mathfrak{R}}$ denotes the element-wise reciprocal of a vector or matrix A . Unlike the Wald and likelihood ratio tests, for a given gene, score tests require fitting the negative binomial model under the null hypothesis only once for all peaks, instead of P full models separately for each peak. This approach facilitates the proposed matrix operations to simultaneously calculate the score tests across all peaks, rather than repeated calculations for one peak at a time. This streamlines computations and significantly enhances computational efficiency.

For studies with finite samples, score tests may be artificially inflated, leading to inaccurate p-values derived from the asymptotic null distribution. To ensure robust inference, we combine an empirical permutation method with genomic control when there is concern about asymptotic results. For gene g , let the original score test statistics across the P peaks be denoted as $\mathbf{SC} = (\mathbf{SC}_1, \dots, \mathbf{SC}_P)$. We then randomly sample the ATAC-Seq data without replacement for L times, and similarly denote the permuted score tests as $\mathbf{SC}^{(l)} = (\mathbf{SC}_1^{(l)}, \dots, \mathbf{SC}_P^{(l)})$, where $l = 1, \dots, L$. Following the genomic control approach, we empirically estimate the inflation factor λ by $\hat{\lambda} = \max(1, \frac{\text{median}(\mathbf{SC}_1^{(1)}, \dots, \mathbf{SC}_P^{(L)})}{0.456})$ and rescale the original score test statistics $\mathbf{SC}_p$ to $\tilde{\mathbf{SC}}_p = \frac{\mathbf{SC}_p}{\hat{\lambda}}$ for $p = 1, \dots, P$, which are then compared to χ_1^2 for p-value calculations.

Results

Simulations

To evaluate *PETScan* in controlling type I error, we conducted simulations with synthetic data as follows. For each gene g ($g = 1, \dots, 100$), we let its expression levels y_{gi} ($i = 1, \dots, N$) follow $NB(\mu_{gi}, \phi)$ with $\log(\mu_{gi}) = x_i$, where the covariate $x_i \sim N(1, 1)$. For each peak p ($p = 1, \dots, 100,000$), its log-transformed chromatin accessibility values were simulated as $z_{pi} \sim N(2, 1)$. We set the dispersion parameter $\phi = 5$ and varied the sample size $N = 50, 100, 200, 500$. We also varied the number of permutations $L = 0, 5, 10, 20$ to investigate the impact of L on type I error control.

We compared the empirical type I error rates and computational efficiency of *PETScan* against the Wald test (Wald) and likelihood ratio test (LRT), two asymptotically equivalent alternatives to the score test, or *PETScan*. Because Wald and LRT require full model parameter estimation, they cannot be efficiently implemented using matrix operations across all P chromatin accessibility peaks, making them computationally prohibitive for large scale genome-wide association analysis. Figure 1 presents the empirical type I error rates of *PETScan* (without permutation and with varying numbers of permutations), Wald and LRT across different sample sizes. As expected, all three tests in general controlled type I error when the sample size was large (e.g., ≥ 200). For smaller sample sizes, the simulation results indicated that genomic control with $L = 5$ permutations was sufficient to control type I error, although slightly conservative. For all real data analyses, we used $L = 20$ to ensure robust results.

In terms of computational efficiency, the runtime of *PETScan* scaled approximately linearly with the number of permutations, consistent with the increasing dimensionality of matrix computations. Overall, *PETScan* was over 1,000 fold faster than both Wald and LRT while achieving comparable performance.

ENCODE human Tissue-Specific data

We applied *PETScan* to paired RNA-Seq and ATAC-Seq data from 46 healthy individuals across 15 human tissues, obtained from the Encyclopedia of DNA Elements (ENCODE) project (ENCODE Project Consortium 2012), which were accessible at <https://www.encodeproject.org/search/?type=Experiment>. The RNA-Seq dataset contained 59,526 genes, and we applied filtering based on gene expression counts and selected variable genes, resulting in a final set of 11,637 genes. Similarly, the ATAC-Seq dataset included 805,169 peaks, which were filtered down to 132,670 peaks. We incorporated gender and the top two principal components derived from log-transformed gene expression data as covariates. P-values culminated in assessing associations across 1,548,656,910 gene-peak pairs. Leveraging parallel computation across 64 CPU cores (AMD EPYC 7313), the analysis was completed in approximately 40 minutes with 20 permutations, significantly outperforming Wald tests, which required over 1,000 times longer. We identified 53,113 significant gene-peak pairs with Bonferroni-adjusted p-values below 0.05, suggesting a widespread impact of chromatin accessibility on gene expression. For further analysis, we focused on the top 7,834 most significant gene-peak pairs with Bonferroni-adjusted p-values below 10^{-5} , as visualized in Figure 2, which was adapted from (Sun 2009) and is available at <http://www.bios.unc.edu/~weisun/software/eMap.pdf>.

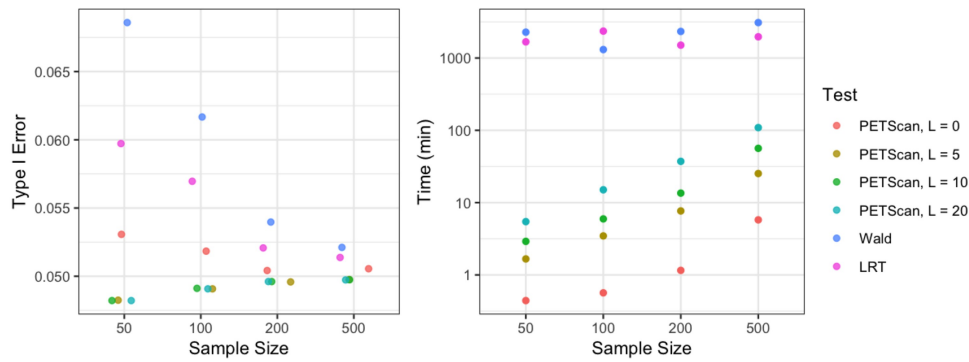


Figure 1. Left: Empirical type I error rates, and Right: Computational time (minutes) for *PETScan* (without permutation, $L = 0$, and with $L = 5, 10, 20$ permutations), Wald and LRT across different sample sizes.

One notable result involved an ATAC-Seq peak at chr12:124,871,497–124,872,709, which was associated with 47 genes across the genome, represented as a vertical line in Figure 2. Among these genes, *HNF1B*, *EGF*, and *SLC4A4* were linked to pathways for maturity onset diabetes of the young, pancreatic cancer, and pancreatic secretion, respectively. As shown in Figure 3, chromatin accessibility at this peak was negatively related to the expression of these associated genes. Notably, the pancreas exhibited lower chromatin accessibility at this peak yet higher gene expression, distinguishing it from other tissues. These findings highlighted the role of chromatin accessibility in tissue-specific gene regulation and shed light on the regulatory landscape of genes involved in pancreatic function and disease. For other associated genes, *BARX2* encoded the transcription factor Barx2, which regulated *GRHL2*; in turn, the transcription factor Grhl2 may potentially target multiple downstream genes, including *BICDL2*, *CA12*, *CLDN4*, *INAVA*, *PKP3*, *RHOV*, and *TRPV6*, as per the Cistrome Data Browser, a comprehensive database for transcriptional and epigenetic regulation studies (Mei et al. 2017, Lambert et al. 2018, Zheng et al. 2019). This regulatory cascade illustrated how chromatin accessibility at a specific peak could influence the coordinated expression of multiple genes across the genome. Such co-association suggested a shared regulatory mechanism, emphasizing the dynamic interplay between chromatin structure and gene regulation.

In addition, *PETScan* identified a few genes whose expression levels were associated with a large number of peaks, represented by horizontal lines in Figure 2. The most prominent was the *RPL32* pseudogene on chromosome 7, which originated from the retrotransposition of *RPL32* mRNA. This association primarily reflected tissue-specific gene expression and chromatin accessibility. Specifically, this gene exhibited higher expression in the pancreas while being lowly expressed in other tissues, while the chromatin accessibility of its associated peaks in the pancreas was either significantly higher or lower compared to other tissues (Figure 4). This finding was consistent with results from the GENCODE project, indicating that the transcription and regulation of pseudogenes were tissue-specific (Pei et al. 2012). A similar pattern was observed for the gene *ST8SIA5* on chromosome 18, which showed significantly higher expression in the adrenal gland, along with chromatin accessibility that was either much higher or lower in this tissue compared to others (Figure 4), suggesting a strong tissue-dependent regulatory mechanism. Existing research has shown that the expression of this

protein-coding gene is linked to open chromatin regions, particularly in brain tissues (Ehrlich et al. 2024). Our study did not include samples from brain tissues, but our analysis suggested that this highly expressed gene in the adrenal gland was also associated with open chromatin regions in the adrenal gland.

To evaluate our findings, we assessed whether significant gene-peak pairs were validated by previously reported cis-eQTLs from the Genotype-Tissue Expression (GTEx) analysis V8 release (Lonsdale et al. 2013), as cis-eQTLs data are more comprehensive than trans-eQTLs data, even though we investigated both cis- and trans-associations. Two gene-peak pairs overlapped with GTEx cis-eQTLs, where variants from significant variant-gene associations were located within peak regions, both exhibiting higher chromatin accessibility and higher gene expression in the pancreas, setting it apart from other tissues (Figure 5). For instance, variant rs4444903, situated within a peak spanning chr4:109,911,610–109,916,356, explained a fraction of the genetic variance in *EGF* expression. Overexpression of *EGF* and its receptor *EGFR* is frequently observed in pancreatic cancer, leading to uncontrolled cell proliferation and tumor progression, and is associated with poor prognosis (Heby et al. 2020). This finding highlighted the tissue-specific accessibility of this region in modulating *EGF* expression, thereby contributing to pancreatic functionality and possibly to the pathological mechanisms underlying pancreatic cancer. Similarly, variant rs66817580 was linked to *FAM3B* expression, which was located within a peak at chr21:41,315,678–41,317,753. *FAM3B*, also known as *PANDER* (pancreatic-derived factor), is primarily expressed in the pancreas, particularly within beta cells. This gene is instrumental in the regulation of glucose metabolism and insulin secretion, and influences beta cell functionality and survival, making it relevant to diabetes and pancreatic disorders (Dong et al. 2023). Further investigation into this regulatory mechanism could provide insights into beta cell dysfunction in diabetes and uncover potential therapeutic targets. *PETScan* demonstrated its effectiveness in uncovering gene regulatory mechanisms, particularly pancreatic-specific associations between chromatin accessibility and targeted gene expression.

10x genomics embryonic mouse brain data

We next conducted *PETScan* on publicly available 10x Genomics embryonic mouse brain data, which included 4,881 cells from the fresh cortex, hippocampus, and ventricular zone of the embryonic mouse brain at day 18 (10x

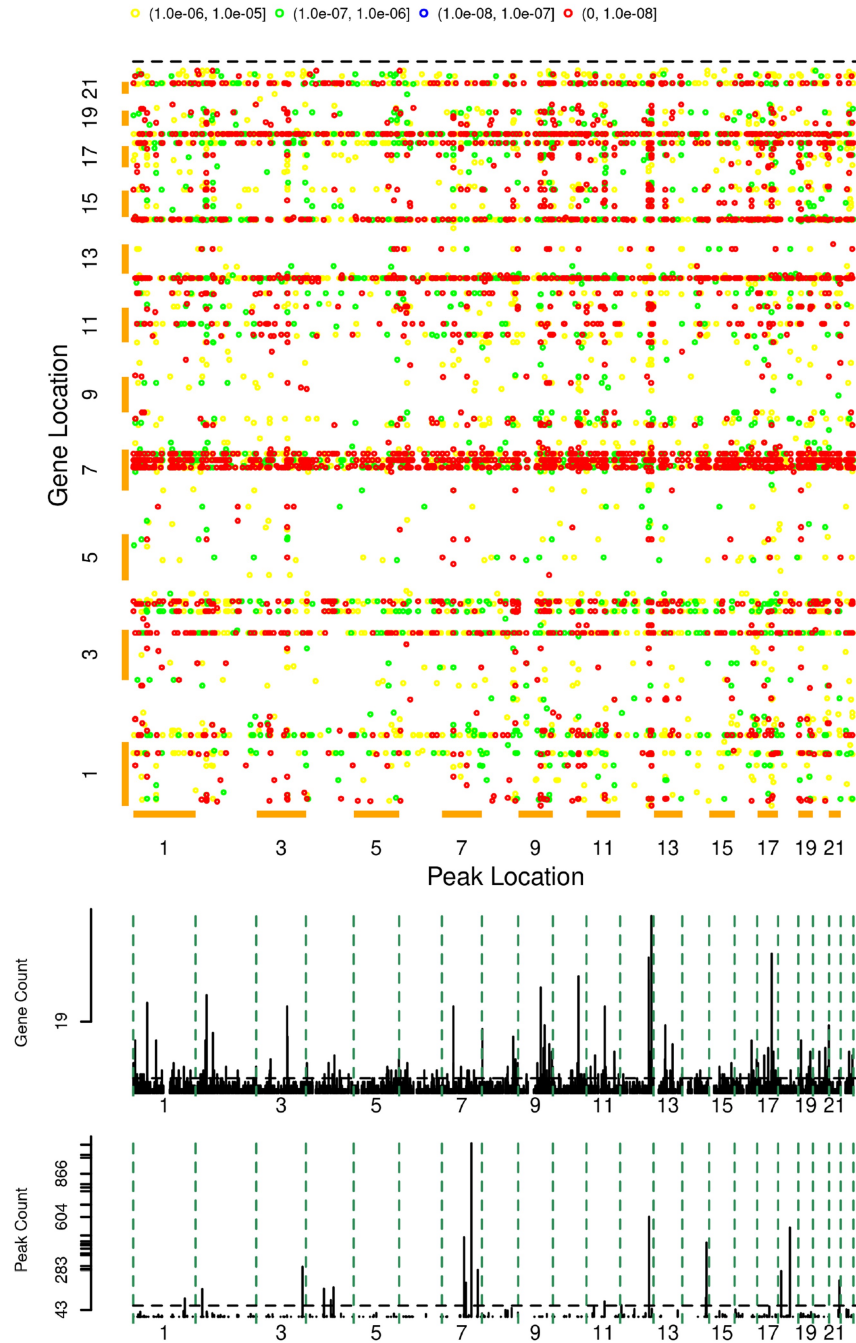


Figure 2. Significant gene-peak pairs across the genome from ENCODE human tissue-specific data are shown, with peak locations on the X-axis and gene locations on the Y-axis. Different levels of significance are indicated using different colors. The middle panel illustrates the number of genes associated with each peak, while the bottom panel displays the number of peaks linked to each gene.

Genomics 2021). The scRNA-Seq and scATAC-Seq data were processed following the TRIPOD vignette (Jiang *et al.* 2022), with cell-type labels transferred from an independent scRNA-Seq reference (La Manno *et al.* 2021) using SAVERCAT (Huang *et al.* 2020). We retained 3,851 cells with consistent transferred labels from four major cell types: neuroblast, forebrain GABAergic neuron, cortical or hippocampal glutamatergic neuron, and glioblast. To address the sparsity issue in single-cell data, we partitioned cells into 84 metacells by clustering based on their scRNA-Seq profiles and assigned cell types to metacells according to the majority cell types within each cluster.

After filtering, we retained 5,935 genes and 42,048 peaks. Using parallel computation across 64 CPU cores, we analyzed 249,554,880 gene-peak pairs in just 7 minutes with 20 permutations, adjusting for the first two principal components of the log-transformed gene expression data. PETScan demonstrated computational efficiency that was three orders of magnitude faster than Wald tests. We identified 26,991 significant gene-peak pairs with Bonferroni-adjusted p-values below 0.05, providing insights into the regulatory landscape of the embryonic mouse brain. For further analysis, we focused on the top 6,858 significant gene-peak pairs with Bonferroni-adjusted p-values below 10^{-5} , as illustrated in Figure 6.

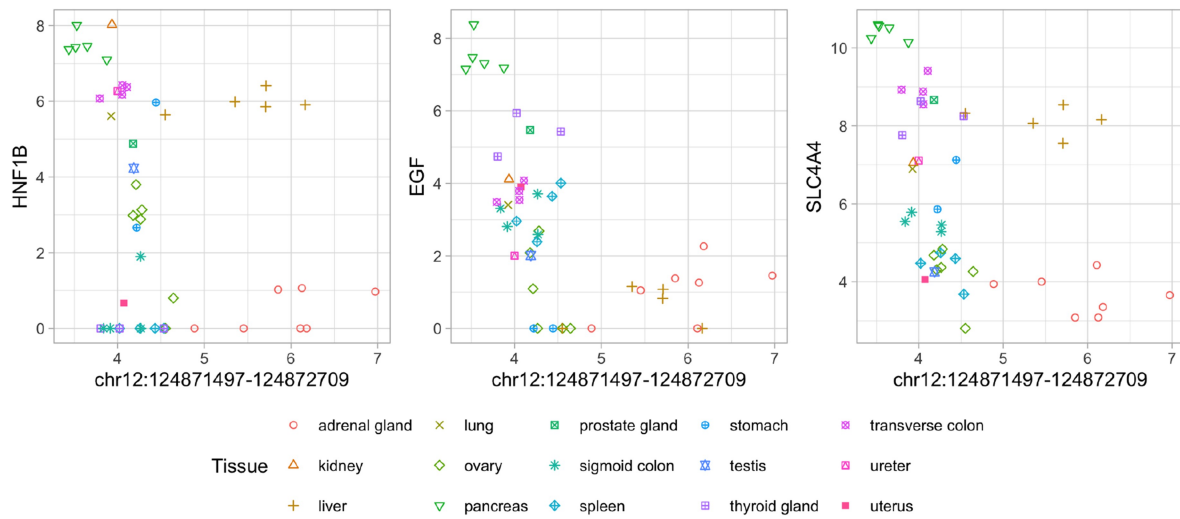


Figure 3. Scatter plots of chromatin accessibility at chr12:124,871,497–124,872,709 and the expression levels of its associated genes.

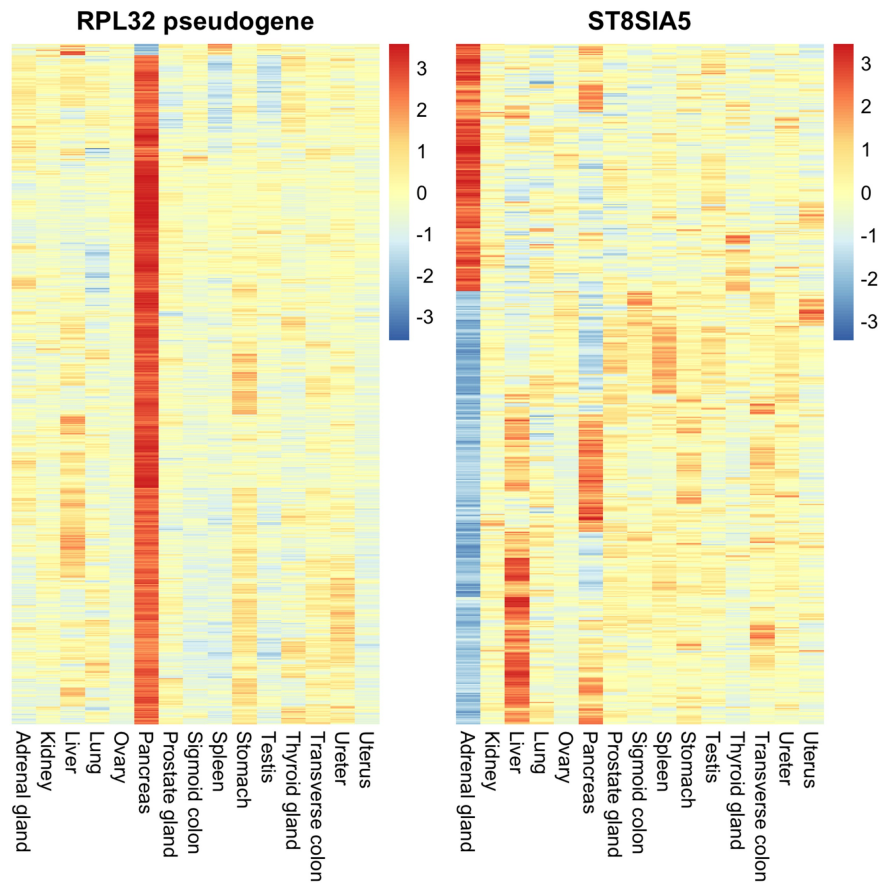


Figure 4. Heatmap of chromatin accessibility of associated peaks for the *RPL32* pseudogene (left) and *ST8SIA5* (right).

We performed Gene Ontology (GO) and KEGG enrichment analyses on genes from these pairs to identify over-represented GO terms and KEGG pathways using clusterProfiler. The top enriched GO terms included fore-brain development, telencephalon development, regulation of neurogenesis, gliogenesis, central nervous system neuron differentiation, and neuron projection guidance. These terms highlighted key processes involved in the development and specialization of the nervous system, particularly regarding

the formation of the brain and its functional components. These findings aligned with the fact that embryonic mouse brain cells were collected during the transition phase between neurogenesis and gliogenesis, further supporting the validity of our results. The top enriched KEGG pathways grouped into two main categories: (1) nervous system: including glutamatergic synapse, GABAergic synapse, and synaptic vesicle cycle, all critical for neurotransmission and communication between neurons; (2) signal transduction: comprising calcium

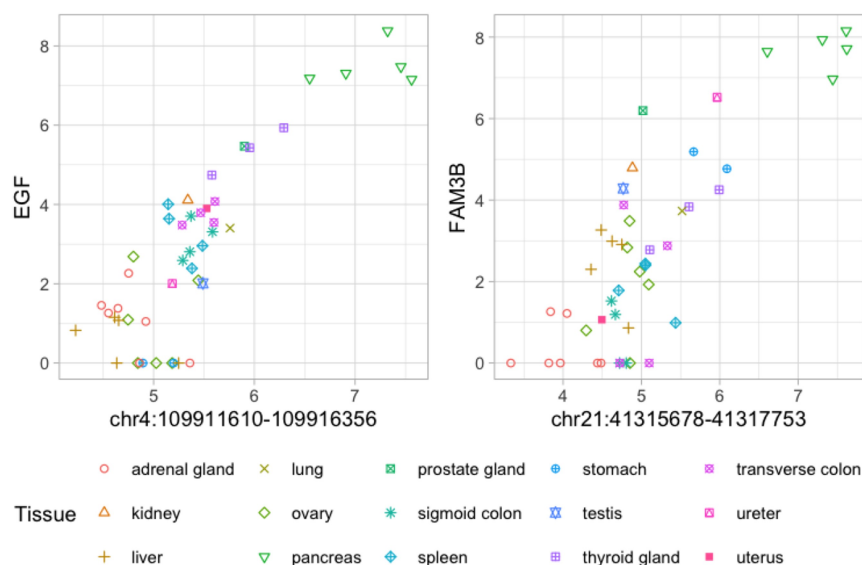


Figure 5. Scatter plots of significant gene-peak pairs from ENCODE human tissue-specific data validated by GTEx cis-eQTLs.

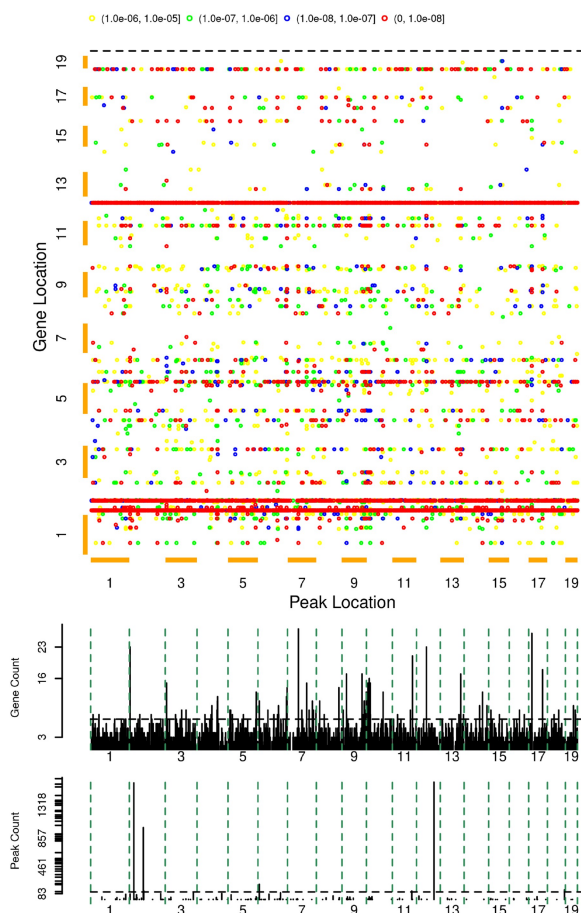


Figure 6. Significant gene-peak pairs across the genome from 10x Genomics embryonic mouse brain data are shown, with peak locations on the X-axis and gene locations on the Y-axis. Different levels of significance are indicated using different colors. The middle panel illustrates the number of genes associated with each peak, while the bottom panel displays the number of peaks linked to each gene.

signaling pathway and neuroactive ligand-receptor interaction, which played essential roles in regulating cellular responses to external stimuli and modulating various

physiological processes. These results were consistent with the characteristics of glutamatergic and GABAergic neurons included in embryonic mouse brain cells.

Furthermore, an ATAC-Seq peak at chr7:54,633,595–54,635,192 was associated with 27 genes across the genome. Notably, open chromatin at this peak was linked to lower expression of *Nrxn3*, *Gad1*, and *Gad2* (Figure 7), all of which were involved in the GABAergic synapse. Conversely, it was correlated with higher expression of *Neurod6*, *Zbtb18*, *Nfix*, *Lhx2*, and *Id2* (Figure 8), all of which were essential for telencephalon development and forebrain development. The upregulated or downregulated genes exhibited functional coherence, indicating that changes in chromatin accessibility may co-regulate genes involved in related biological processes.

To validate our findings, we compared significant gene-peak pairs with previously reported chromatin interactions, using PLAC-Seq data of the mouse fetal forebrain (Zhu *et al.* 2019). Our analysis identified nine significant gene-peak pairs in PLAC-Seq data, all exhibiting positive correlations between chromatin accessibility and gene expression, predominantly in GABAergic and glutamatergic neurons (Figure 9). We also performed *PETScan* to evaluate cis-associations for all genes and their corresponding peaks within 1,000 kb of the transcription start site, identifying 4,211 significant gene-peak pairs with Benjamini & Hochberg (BH) adjusted p-values below 0.05, of which 756 (18.0%) were validated by PLAC-Seq data. These results showed the biological relevance of our approach in capturing chromatin interactions that regulate gene expression in mouse brain development.

Discussion

PETScan is a computationally efficient tool for providing a comprehensive perspective on tissue- or cell-type-specific gene regulation and its implications for diverse biological processes and diseases. Our findings align with previously reported cis-eQTLs or PLAC-Seq data, validating the robustness and biological relevance of the identified gene-peak pairs. This concordance underscores *PETScan*'s potential to

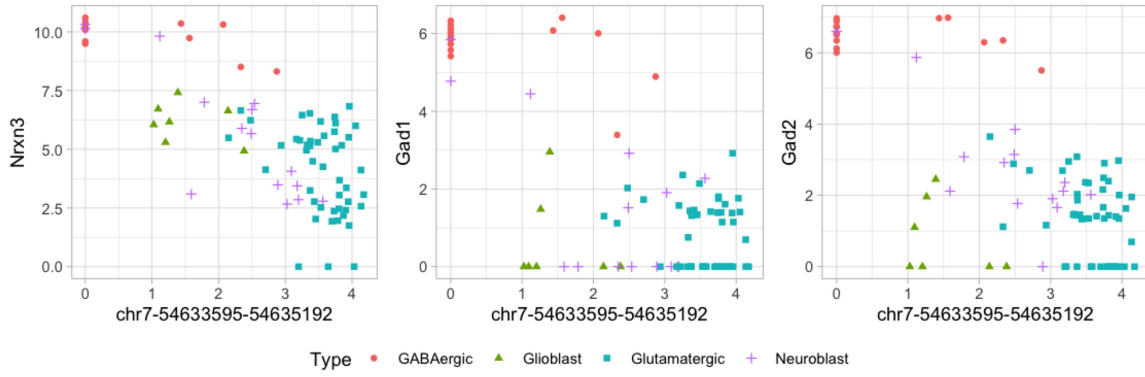


Figure 7. Scatter plots of chromatin accessibility at chr7:54,633,595–54,635,192 and the expression levels of its negatively associated genes.

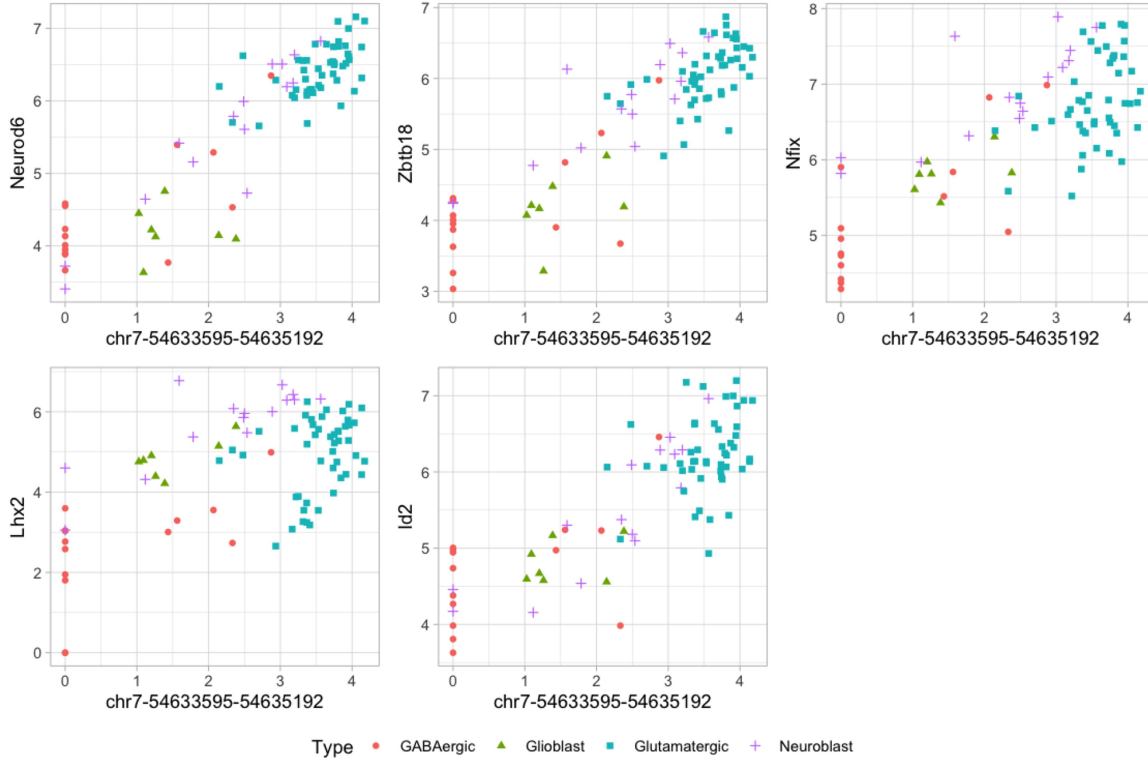


Figure 8. Scatter plots of chromatin accessibility at chr7:54,633,595–54,635,192 and the expression levels of its positively associated genes.

effectively bridge chromatin accessibility and gene expression data.

Beyond cis-associations, *PETScan* identifies trans-associations, capturing interactions between regulatory elements and distal target genes to uncover long-range regulatory mechanisms that may not be detectable using traditional approaches. Genome-wide associations identified by *PETScan* offer important insights into regulatory mechanisms, providing experimental validation guidelines for confirming the functional roles of specific open chromatin regions in regulating the transcriptomic expressions of their targeted genes.

Although *PETScan* is developed for performing billions of association tests between high-dimensional RNA-Seq and ATAC-Seq data, it is also versatile for other paired omics data, provided that negative binomial models offer an appropriate framework. For example, *PETScan* can be a viable approach for eQTL analysis when gene expression profiles are

measured using RNA-Seq data, where the normality assumption of Matrix eQTL is no longer appropriate. Additionally, *PETScan* can facilitate genome-wide association analysis between RNA-Seq and ChIP-Seq data, evaluating how transcription factor binding or histone modifications impact gene expression. Similarly, analysis of paired RNA-Seq and Bisulfite-Seq data with *PETScan* can provide insights into how DNA methylation regulates gene expression.

Less obviously, *PETScan* can be adopted for differential gene expression analysis of RNA-Seq with a given factor, such as a binary outcome. When the sample size is limited, necessitating a large number of permutations, say P , to ensure valid type I error control, the P sets of permuted outcomes can be treated analogously to chromatin accessibility measurements across P peaks, on which *PETScan* can be applied to derive an empirical null distribution. Genomic control can similarly be implemented through permutations to address potential limitations associated with small sample sizes,

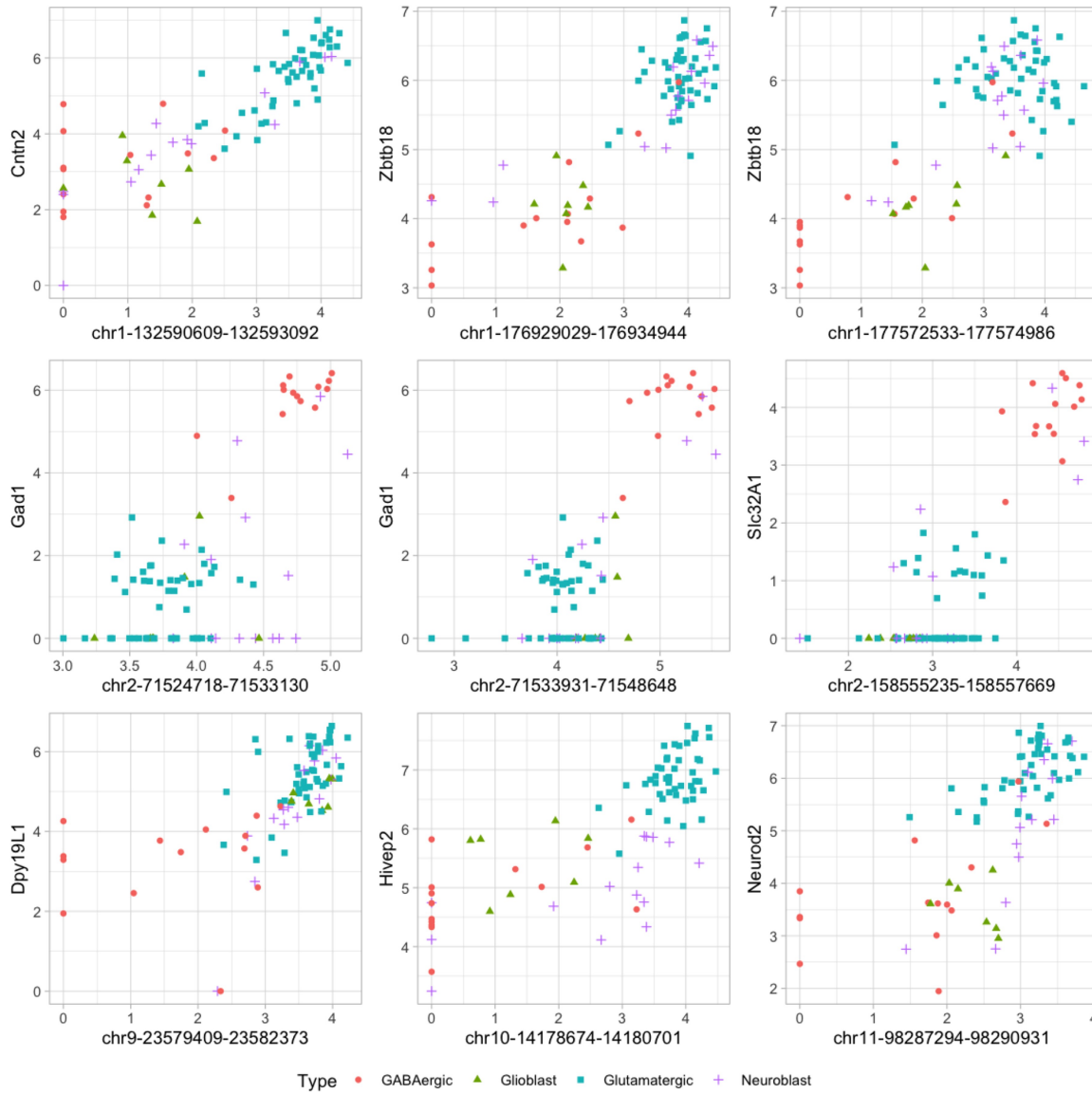


Figure 9. Scatter plots of significant gene-peak pairs from 10x Genomics embryonic mouse brain data validated by PLAC-Seq data.

thereby enhancing the reliability and robustness of the association findings. Similar research direction can be found in (Barry *et al.* 2025).

A notable feature of *PETScan* is its reliance on score tests, which focus solely on assessing the presence of associations without estimating effect sizes and directionality (positive or negative) as traditional regression-based methods do. While this approach offers computational efficiency and stability, it limits the ability to determine whether a peak positively or negatively influences gene expression, a crucial factor in elucidating underlying biological mechanisms. Nevertheless, *PETScan* remains a powerful tool for large-scale screening in genome-wide association analysis, particularly when computational efficiency is paramount. By identifying candidate interactions for further validation, *PETScan* provides a foundation for subsequent in-depth investigations.

The use of negative binomial models provides the advantage of flexible covariate adjustment compared to simple correlations. Currently, *PETScan* applies the same set of covariates to all genes. However, some covariates may influence the relationship between chromatin accessibility and

gene expression in a gene-specific manner, such as the cis-effect of DNA methylation and transcript factor expression levels. *PETScan* can be easily modified to include gene-specific covariates when relevant data is available, enabling more accurate inference.

When multiple paired RNA-Seq and ATAC-Seq samples are collected from the same subject, such as from different biological time points, the correlations between these samples must be accounted for. Our future plan is to extend *PETScan* from negative binomial models to negative binomial mixed models (Tsonaka and Spitali 2021, Sun *et al.* 2016, Rivellese *et al.* 2022), presenting an even more computationally daunting task. In addition, negative binomial models in *PETScan* may not adequately account for counts with excess zeros commonly observed in other omics data, such as single-cell RNA-Seq data and microbiome data, suggesting the need for zero-inflated Poisson or zero-inflated negative binomial models, which could be another useful extension direction for *PETScan*.

In conclusion, *PETScan* offers a computationally efficient and biologically meaningful framework for assessing

genome-wide associations between chromatin accessibility and gene expression. It holds promise for advancing the study of gene regulation and guiding experimental validation efforts, ultimately contributing to a deeper understanding of the regulatory mechanisms underlying gene expression.

Author contributions

Yajing Hao (Formal analysis [Equal], Methodology [Equal], Software [Equal], Visualization [Equal], Writing—original draft [Equal]), and Tal Kafri (Funding acquisition [Equal], Resources [Equal], Supervision [Equal], Writing—review & editing [Equal]), Fei Zou (Conceptualization [Equal], Funding acquisition [Equal], Methodology [Equal], Supervision [Equal], Writing—review & editing [Equal]).

Conflict of interest: None declared.

Funding

The study is supported by the National Institutes of Health [R01-HL155986].

References

- 10x Genomics 2021. Fresh Embryonic E18 Mouse Brain (5k). Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger ARC 2.0.0.
- Ackermann AM, Wang Z, Schug J *et al.* Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol Metab* 2016;5:233–44.
- Bai Y, Deng X, Chen D *et al.* Integrative analysis based on ATAC-seq and RNA-seq reveals a novel oncogene PRPF3 in hepatocellular carcinoma. *Clin Epigenetics* 2024;16:154.
- Barry T *et al.* 2025. The permuted score test for robust differential expression analysis. *arXiv*, <https://doi.org/10.48550/arXiv.2501.03530>, preprint: not peer reviewed.
- Bravo González-Blas C, De Winter S, Hulselmanns G *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;20:1355–67.
- Buenrostro JD, Giresi PG, Zaba LC *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 2013;10:1213–8.
- Cao J, Cusanovich DA, Ramani V *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 2018;361:1380–5.
- Casamassimi A, Federico A, Rienzo M *et al.* Transcriptome profiling in human diseases: new advances and perspectives. *Int J Mol Sci* 2017;18:1652.
- Chen S, Lake BB, Zhang K *et al.* High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* 2019;37:1452–7.
- Dai W, Qiao X, Fang Y *et al.* Epigenetics-targeted drugs: current paradigms and future challenges. *Signal Transduct Target Ther* 2024;9:332.
- Dean A. In the loop: long range chromatin interactions and gene regulation. *Brief Funct Genomics* 2011;10:3–10.
- Dekker J, Misteli T. Long-range chromatin interactions. *Cold Spring Harb Perspect Biol* 2015;7:a019356.
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- Dhar GA, Saha S, Mitra P *et al.* DNA methylation and regulation of gene expression: guardian of our health. *Nucleus (Calcutta)* 2021;64:259–70.
- Di Y, Schafer DW, Cumbie JS *et al.* The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Statistical Applications in Genetics and Molecular Biology* 2011;10:1–28.
- Ding M, Huang W, Liu G *et al.* Integration of ATAC-seq and RNA-seq reveals FOSL2 drives human liver progenitor-like cell aging by regulating inflammatory factors. *BMC Genomics* 2023;24:260.
- Dong Q-T, Ma D-D, Gong Q *et al.* FAM3 family genes are associated with prognostic value of human cancer: a pan-cancer analysis. *Sci Rep* 2023;13:15144.
- Ehrlich M, Ehrlich KC, Lacey M *et al.* Epigenetics of genes preferentially expressed in dissimilar cell populations: myoblasts and cerebellum. *Epigenomes* 2024;8:4.
- ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.
- Frankel N. Multiple layers of complexity in *cis*-regulatory regions of developmental genes. *Dev Dyn* 2012;241:1857–66.
- Gagnidze K, Pfaff DW. 2022. Epigenetic mechanisms: DNA methylation and histone protein modification. In: *Neuroscience in the 21st Century*. Cham: Springer International Publishing, 2677–716.
- Heby M, Karnevi E, Elebro J *et al.* Additive clinical impact of epidermal growth factor receptor and podocalyxin-like protein expression in pancreatic and periampullary adenocarcinomas. *Sci Rep* 2020;10:10373.
- Huang M *et al.* 2020. Dimension reduction and denoising of single-cell RNA sequencing data in the presence of observed confounding variables. *bioRxiv*. <https://doi.org/10.1101/2020.08.03.234765>
- Huisinga KL, Brower-Toland B, Elgin SCR *et al.* The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma* 2006;115:110–22.
- Jiang Y, Harigaya Y, Zhang Z *et al.* Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and *cis*-regulatory regions. *Cell Syst* 2022;13:737–51.e4.
- Kadauke S, Blobel GA. Chromatin loops in gene regulation. *Biochim Biophys Acta* 2009;1789:17–25.
- Kartha VK, Duarte FM, Hu Y *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genom* 2022;2:100166.
- Klemm SL, Shipony Z, Greenleaf WJ *et al.* Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 2019;20:207–20.
- La Manno G, Siletti K, Furlan A *et al.* Molecular architecture of the developing mouse brain. *Nature* 2021;596:92–6.
- Lambert SA, Jolma A, Campitelli LF *et al.* The human transcription factors. *Cell* 2018;172:650–65.
- Ledru N, Wilson PC, Muto Y *et al.* Predicting proximal tubule failed repair drivers through regularized regression analysis of single cell multiomic sequencing. *Nat Commun* 2024;15:1291.
- Lohia R, Fox N, Gillis J *et al.* A global high-density chromatin interaction network reveals functional long-range and trans-chromosomal relationships. *Genome Biol* 2022;23:238.
- Lonsdale J, Thomas J, Salvatore M *et al.* The Genotype-Tissue expression (GTEx) project. *Nat Genet* 2013;45:580–5.
- Love MI, Huber W, Anders S *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
- Ma S, Zhang B, LaFave LM *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020;183:1103–16.e20.
- Maass PG, Barutcu AR, Rinn JL *et al.* Interchromosomal interactions: a genomic love story of kissing chromosomes. *J Cell Biol* 2019;218:27–38.
- Mei S, Qin Q, Wu Q *et al.* Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res* 2017;45:D658–D662.
- Miele A, Dekker J. Long-range chromosomal interactions and gene regulation. *Mol Biosyst* 2008;4:1046–57.
- Nagalakshmi U, Wang Z, Waern K *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;320:1344–9.
- Park J-W, Rhee J-K. Integrative analysis of ATAC-seq and RNA-seq through machine learning identifies 10 signature genes for breast cancer intrinsic subtypes. *Biology (Basel)* 2024;13:799.
- Pei B, Sisu C, Frankish A *et al.* The GENCODE pseudogene resource. *Genome Biol* 2012;13:R51.
- Ralston A, Shaw K. Gene expression regulates cell differentiation. *Nature Education* 2008;1:127.

- Rivellese F, Surace AEA, Goldmann K, R4RA collaborative group *et al.* Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nat Med* 2022;28:1256–68.
- Kundaje A, Meuleman W, Ernst J, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30.
- Robinson MD, McCarthy DJ, Smyth GK *et al.* edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40.
- Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 2012;28:1353–8.
- Song Q, Hou Y, Zhang Y *et al.* Integrated multi-omics approach revealed cellular senescence landscape. *Nucleic Acids Res* 2022; 50:10947–63.
- Staitieh BS, Hu X, Yeligar SM *et al.* Paired ATAC- and RNA-seq offer insight into the impact of HIV on alveolar macrophages: a pilot study. *Sci Rep* 2023;13:15276.
- Starks RR, Biswas A, Jain A *et al.* Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin* 2019;12:16.
- Stuart T, Srivastava A, Madad S *et al.* Single-cell chromatin state analysis with signac. *Nat Methods* 2021;18:1333–41.
- Sun W. 2009. eQTL analysis by Linear Model. <http://www.bios.unc.edu/~weisun/software/eMap.pdf>
- Sun X, Dalpiaz D, Wu D *et al.* Statistical inference for time course RNA-Seq data using a negative binomial mixed-effect model. *BMC Bioinformatics* 2016;17:324.
- Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014;7:33.
- Tsonaka R, Spitali P. Negative binomial mixed models estimated with the maximum likelihood method can be used for longitudinal RNAseq data. *Brief Bioinform* 2021;22:bbaa264.
- van den Heuvel A, Stadhouders R, Andrieu-Soler C *et al.* Long-range gene regulation and novel therapeutic applications. *Blood* 2015; 125:1521–5.
- Yang H, Li G, Qiu G *et al.* Bioinformatics analysis using ATAC-seq and RNA-seq for the identification of 15 gene signatures associated with the prediction of prognosis in hepatocellular carcinoma. *Front Oncol* 2021;11:726551.
- Zeng H. What is a cell type and how to define it? *Cell* 2022; 185:2739–55.
- Zhang Y, Jiang M, Xiong Y *et al.* Integrated analysis of ATAC-seq and RNA-seq unveils the role of ferroptosis in PM2.5-induced asthma exacerbation. *Int Immunopharmacol* 2023;125:111209.
- Zheng R, Wan C, Mei S *et al.* Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* 2019;47:D729–35.
- Zhu C, Yu M, Huang H *et al.* An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* 2019;26:1063–70.
- Zhu T, Zhou X, You Y *et al.* cisDyNet: an integrated platform for modeling gene-regulatory dynamics and networks. *Imeta* 2023;2:e152.
- Zou F, Sun W, Crowley JJ *et al.* A novel statistical approach for jointly analyzing Rna-Seq data from F1 reciprocal crosses and inbred lines. *Genetics* 2014;197:389–99.