Identifying dysregulated regions in amyotrophic lateral sclerosis through DNA accessibility outliers

Muhammed Hasan Celik[1,2], Julien Gagneur[3,4,5,6], Ryan G Lim[7], Jie Wu[8], Leslie M. Thompson[7,8,9,10,11], Xiaohui Xie[1]

1 Department of Computer Science, University of California Irvine, Irvine, CA, USA
2 Center for Complex Biological Systems, University of California Irvine, Irvine, CA, USA
3 Department of Informatics, Technical University of Munich, Garching, Germany
4 Helmholtz Association - Munich School for Data Science (MUDS), Munich, Germany
5 Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany
6 Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany
7 Institute for Memory Impairments and Neurological Disorders, University of California Irvine, Irvine, CA 92697, USA
8 Department of Biological Chemistry, University of California Irvine, Irvine, CA, USA
9 UCI MIND, University of California Irvine, Irvine, CA, USA
10 Department of Psychiatry and Human Behavior and Sue and Bill Gross Stem Cell Center, University of California Irvine, Irvine, CA, USA
11 Department of Neurobiology and Behavior, University of California Irvine, Irvine, CA, USA

Abstract

The high heritability of ALS and similar rare diseases contrasts with their low molecular diagnosis rate post-genetic testing, pointing to potential undiscovered genetic factors. DNA accessibility assays quantify the activity of functional elements genome-wide, offering invaluable insights into dysregulated regions. In this research, we introduced EpiOut, a computational toolbox to identify outliers in DNA accessibility. These outliers represent dysregulated regions where DNA accessibility uniquely diverges from the population baseline in a single or few samples. Annotation of accessible regions with histone ChIP-seq and Hi-C indicates that outliers are concentrated in functional loci, especially among promoters interacting with active enhancers. Across different omics levels, outliers are robustly replicated, and DNA accessibility outliers are reliable predictors of gene expression outliers and aberrant protein levels. For example, 59% of gene expression outliers can be linked to epigenetic changes. A comparison of promoter activity against gene expression reveals if aberration in the molecule phenotype is linked to pre-transcriptional or post-transcriptional regulation. Our findings demonstrate that the outlier detection paradigm can uncover dysregulated regions in rare diseases. EpiOut is open-sourced and freely available at github.com/uci-cbcl/EpiOut.

Introduction

Amyotrophic lateral sclerosis (ALS) is a rare neuromuscular degenerative disease affecting 0.6 to 3.8 per 100,000 people with a poor survival prognosis without a cure[1,2]. ALS is a complex disease where a single gene or pathway cannot explain the disease phenotype due to the heterogeneity of genetic causes and over 30 genes associated with ALS[1,3]. Meta-analysis and twin studies estimate the heritability of ALS disease at 61% (with 38%-78% confidence intervals) in sporadic cases (sALS), patients without a history of the disease in the family[4]. Despite the high heritability of ALS, 11 to 21% of patients[5,6] receive a diagnosis after genetic testing. The gap between high heritability and low diagnostic rate implies the existence of many undiscovered ALS-related genes.

There are large-scale sequencing efforts to discover the genetic bases of ALS[7,8]. These studies utilized GWAS, QTL, and differential expression analysis from a large cohort of samples to detect aberrations in ALS patients compared to control samples[7,9–11]. These statistical approaches successfully detected the most common factors (variants, genes, and pathways) associated with disease phenotype, yet detecting rare genetic factors remains challenging due to low statistical power. Outlier detection is a complementary statistical approach that uncovers aberrations specific to one or a few patients. Applying the outlier detection paradigm to transcriptomics data has revealed the dysregulation of many novel splicing and gene expression outliers[12–18]. This approach offers a promising direction for enhancing diagnostic rates, as it effectively models the heterogeneous nature of rare diseases.

The outlier detection approach has recently been applied to proteomics, and robust replication of aberrations across multiple omics data demonstrates the reliability of the outliers for disease diagnostics[19]. Expanding the outlier detection approach to epigenetics could provide further insight into the dysregulation of functional regions and their impact on gene expression in disease. Because epigenetic mechanisms, such as chromatin accessibility and enhancer/promoter activity, control gene expression. Despite the widespread use of chromatin accessibility assays such as ATAC-seq and DNase-seq enabling genome-wide investigation of the epigenetic landscape[20], the outlier detection approach has not been applied to chromatin accessibility data to the best of our knowledge.

Controlling for both known and latent confounding factors is crucial to exclude outliers resulting from technical artifacts or confounding factors during outlier detection. A linear autoencoder (LR-AE) is an effective statistical method for obtaining latent confounders from high-dimensional data. For example, PCA is a specific case of an LR-AE that can detect and correct for latent confounders better than alternative statistical methods[21], such as PEER[22]. However, PCA has its limitations as well. PCA minimizes the Euclidean distance between measured and reconstructed counts; thus, it is suboptimal for discrete read counts of omics data with high dispersion and uncertainty due to low-coverage[23].

The LR-AE with negative binomial objective function proposed by OUTRIDER[24] is better suited for read counts and has superior outlier detection performance. However, this approach is much more computationally expensive to optimize than PCA. OUTRIDER can be applied to transcriptomics data with a dimensionality ranging from ten to fifteen thousand expressed protein-coding genes. However, the LR-AE implementation of OUTRIDER may not scale to DNA accessibility with a dimensionality of hundreds of thousands of genomic regions. Nevertheless, DNA accessibility outlier detection with an LR-AE has not been previously explored.

Here, we present EpiOut, a software developed for DNA accessibility outlier detection. Our proposed method takes read alignment files and accessible regions as input, performs ultra-fast read counting per accessible region, detects outliers using an LR-AE with a negative binomial objective function, and annotates outlier regions using ChIP-seq and Hi-C. Optimization of the decoder layer and dispersion parameters requires solving a large number of independent convex problems. We significantly accelerated the LR-AE using TensorFlow[25] by utilizing a vectorized variation of backtracking line search for the dispersion parameters and L-BFGS for the decoder layer.

We applied EpiOut to DNA accessibility data from ALS samples to identify epigenetic aberrations that might be linked to the disease phenotype. EpiOut pinpoints a small number of sample-specific loci as outliers. These outliers are significantly enriched for functional regions as indicated by histone ChIP-seq experiments, and these outlier regions potentially interact with each other based on the Hi-C data. The accessibility outliers are reliable predictors of outliers detected in transcriptomic and proteomic analyses. Moreover, we observed a substantial depletion of high-impact rare coding variants in gene expression outliers when the promoters of these genes are accessibility outliers. In contrast, the enrichment of high-impact variants was detected in cases where their promoters were not outliers. Our results demonstrate that comparing epigenetic and transcriptomic outliers can offer valuable insights into whether aberrations in gene expression are influenced by pre-transcriptional or post-transcriptional regulation.

Results

In this study, we explored aberrant DNA accessibility in ALS using ATAC-seq experiments from the AnswerALS dataset (Methods). This dataset comprises paired genomics, transcriptomics, accessibility, and proteomics experiments from 253 individuals. With our novel method, EpiOut, we pinpointed regions with abnormal DNA accessibility and investigated the molecular impact of these outliers by comparing paired experiments across different omics levels. Our results highlight the biological relevance and robustness of the identified accessibility outliers.

102 **Detection of accessible regions**

103 Detection of aberrant accessibility requires read counts for a set of accessible regions consistent
104 across the individuals. First, we merged ATAC-seq reads of all samples into one meta-sample, then
105 performed joint peak calling with MACS2[26] using the meta-sample to obtain a consistent set of
106 accessible regions. We detected a total of $858{,}268$ peaks before any filtering. Next, we counted the
107 number of ATAC-seq reads overlapping with each accessibility region. The counting read from a large
108 sample cohort is computationally intensive[14]. Thus, we developed EpiCount, an efficient read counter
109 designed for accessibility data. EpiCount is twice as fast as the state-of-the-art counting tool[27]
110 (Methods), using ~ 50 times less memory (Extended Data Fig. 1a,b) and fifteen times faster than the
111 reported runtime of commonly used counting methods[14]. The large memory footprint of the counting
112 method is a limiting factor for the embarrassing parallelization of counting across the samples; thus,
113 EpiCount is highly parallelizable across multiple CPU cores. Lastly, we filtered accessible regions based
114 on the read counts because many accessible regions often do not replicate across samples. We
115 imposed a replication filter to these peaks, ensuring that the accessible regions were observed in at
116 least 50% of the samples with a minimum of two reads and exhibited high accessibility (100 reads) in
117 at least one sample (Extended Data Fig. 2a). Applying these filters yielded $114{,}428$ accessible regions
118 replicated across samples.

119 **Outlier detection and benchmark**

120 We hypothesize that major covariation between samples is primarily due to confounding factors.
121 Therefore, eliminating this major covariation is essential in the proposed methodology to spotlight
122 the rare aberrations potentially associated with the disease phenotype. To substantiate this premise,
123 we examined the relationships between the principal components of accessibility data and the
124 disease status of the samples. Fig. 1a illustrates the lack of clustering between samples by disease
125 phenotype using the top two principal components of DNA accessibility. The top 25 principal
126 components account for approximately 94% of the DNA accessibility variation between samples.
127 However, none of these principal components show a significant association with the disease
128 phenotype (Extended Data Fig. 3a,b). These findings indicate that major covariations in accessibility
129 are likely attributable to confounding factors.

130 To evaluate the performance of the outlier detection methods, we employed an artificial outlier
131 injection procedure previously proposed for detecting aberrant gene expression[24]. To create ground
132 truth, we injected large aberrations, called artificial outliers, to read counts of ALS samples and then
133 benchmarked the performance of tools to classify those artificial outliers on the area under the
134 precision-recall curve (Methods) (Fig. 1c). In naive negative binomial, we estimated the mean of the

negative binomial test as a sample mean of read count per ATAC-seq peaks (Methods) and estimated dispersion with maximum likelihood, then ranked predictions by p-value based on the negative binomial test. The naive negative binomial model performs poorly with auPRC of $1.8 \pm 1\%$ because the expected read counts of the naive negative binomial model are not sample-specific. As an alternative outlier detection method, expected read counts per peak and sample can be estimated by PCA (Methods), and ranking predicted outliers by Z-score based on the expected and observed read counts of the PCA model have the performance of auPRC $9.7 \pm 0.9\%$. Our proposed method estimates expected read counts using an LR-AE (Fig. 1b). Expected read counts are incorporated into a negative binomial test as the mean parameter, and the dispersion is estimated based on the expected and observed counts (Methods). In our implementation of the LR-AE, the weights of the encoder and decoder layers are initialized with the rotation matrix of PCA. Then, the dispersion parameter is initially estimated using expected counts based on initial weights, and weights of the decoder layer are updated to maximize negative binomial likelihood using initial dispersion estimation. After the decoder layer optimization, we recalculate the dispersion estimation and apply the negative binomial test to estimate outliers. Also, it is critical to account for read coverage differences between samples (Extend Data Fig. 2b); thus, read counts are normalized for size factors[28] (Methods). We chose the optimal bottleneck size of LR-AE with hyperparameter tuning on the validation set (Extended Data Fig. 4). This approach outperforms the previous two methods by achieving an auPRC of $50.3 \pm 0.2\%$. An alternative LR-AE-based method, OUTRIDER, outperforms PCA and performs similarly with EpiOut (auPRC of $48 \pm 0.2\%$). Both methods employ an LR-AE with a negative binomial, a more proper distribution to fit counts data than PCA (Methods). Those results show that the estimation of dispersion and gene/sample-specific accessibility expectation followed by the negative binomial test is critical for outlier detection.

DNA accessibility is higher dimensional than gene expression because accessibility data may contain hundreds of thousands of accessible genomic regions. In contrast, gene expression data only contains around ten to fifteen thousand expressed protein-coding genes. Thus, the scalability of the outlier detection approach is essential to apply the method to high dimensional DNA accessibility data. Although OUTRIDER and EpiOut have similar auPRC scores, the autoencoder implementation of OUTRIDER is significantly slower than our proposed method (Fig. 1d). For example, outlier detection with EpiOut (745 ± 49 seconds) is 60 times faster than OUTRIDER ($44{,}874 \pm 200$ seconds). Our implementation is faster due to a couple of reasons. Firstly, we do not optimize the encoder layer of the autoencoder because we observed that the initial estimation of encoder weights with PCA is close to optimal, so further training of the encoder is unnecessary. Also, EpiOut only performs one alternating optimization step to fit the decoder layer and estimate dispersion. In contrast, OUTRIDER

169 performs multiple alternative optimization steps to estimate dispersion and train the encoder and
170 decoder layers. Lastly, the optimization of the decoder layer and the estimation of dispersion
171 parameters necessitate solving a multitude of independent convex optimization problems. We
172 efficiently approach these using a vectorized L-BFGS for the decoder layer and a vectorized and
173 bounded backtracking line search for dispersion estimation, all implemented in TensorFlow
174 (Methods).

175 We further integrated known confounding factors such as batch ID, sex, reported race, and ethnicity
176 into outlier detection (Methods). The contribution of each feature for outlier detection is summarized
177 in Fig. 1e. The naive negative binomial test estimates 76.5 ± 614 outlier regions per sample, and the
178 inclusion of latent confounding factors for expected read count estimation significantly reduces the
179 number of outliers to 18.3 ± 24. Known confounding factors such as batch ID, sex, race, and ethnicity
180 further reduce the number of an outlier to 14.2 ± 13.6 out of $114{,}428$ accessible regions per sample.
181 Thus, our method pinpoints only a handful of aberrant regions per sample.

182 As another benchmark, we compared cross-correlation between samples. Samples are highly cross-
183 correlated based on the raw read counts (Fig. 1f), and samples do not cluster by the disease phenotype
184 but rather by potential latent cofounders. Correcting for latent and known confounding factors and
185 normalization of read counts decorrelate samples and eliminate clusters due to potential batch effects
186 (Fig. 1g). Samples have an average cross-correlation of 96.1% based on the raw read counts. In
187 comparison, there is a 0% average cross-correlation after the proposed correction method ($\mathrm{P} = 0$
188 based on the paired Wilcoxon test, Fig. 1h).

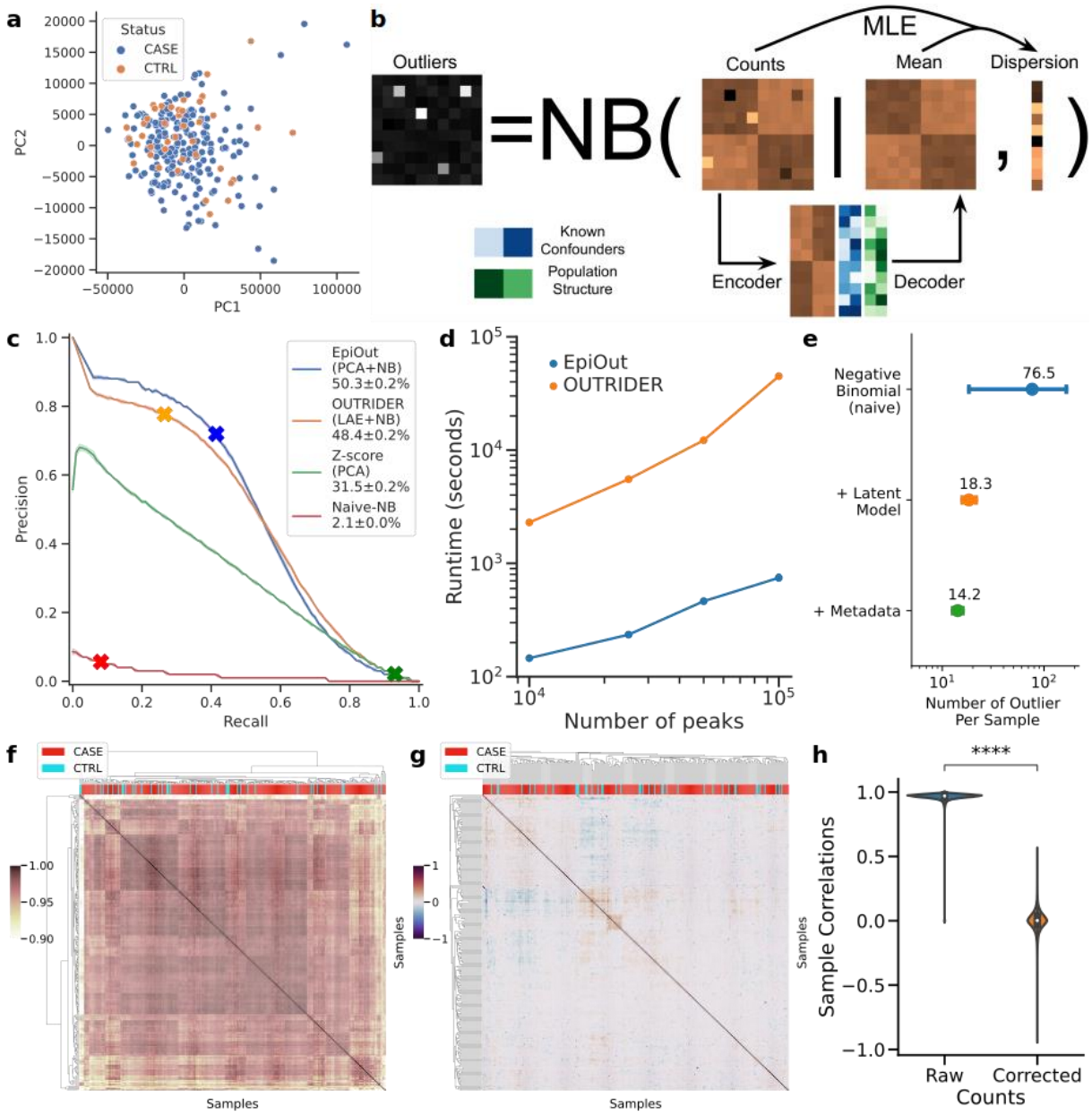

Fig. 1: Proposed outlier detection methodology and its benchmark (a) Samples do not form distinct clusters based on their disease phenotype according to the top two principal components of DNA accessibility. **(b)** The architecture of the proposed methodology for outlier detection (EpiOut). The approach employs the negative binomial test. The mean parameter of the negative binomial distribution is predicted with an LR-AE, which uses latent confounders obtained from data in addition to reported known confounding as features to predict sample-specific expected accessibility. The dispersion parameter is fitted with MLE using the observed and expected counts. **(c)** A precision-recall curve shows the performance of alternative outlier detection methods. Methods were benchmarked based on the classification accuracy of the injected artificial outlier. Specific cutoffs of models are indicated with cross marks (an absolute Z-score of 2 for PCA, a p-value of 0.05 for naive negative binomial, OUTRIDER, and EpiOut). **(d)** Runtime benchmark of outlier detection methods. **(e)** Contribution of each component (such as latent and known confounding factors) of the model to reduce the number of outliers per sample. **(f)** A Cluster heatmap displaying the cross-correlation of samples based on the raw accessibility counts and **(g)** cross-correlation of samples after correction of

counts. **(h)** Distribution of cross-correlation between sample pairs before and after correction of accessibility reads.

Functional annotation of accessible regions and outliers

The biological relevance of accessibility outliers remains ambiguous unless annotated with epigenetic marks. Thus, we developed EpiAnnot, which annotated accessible regions for histone marks and 3D chromatin interactions based on ChIP-seq and Hi-C experiments. EpiAnnot integrates publicly available ChIP-seq and Hi-C for cell lines/tissues available in Roadmap Epigenomics[29] and ENCODE[30] or from custom data sources. We annotated the previously identified accessible regions and outliers using the H3K4me3, H3K27ac, and H3K4me1 histone marks observed in motor neurons. These neurons were derived from the iPSCs of clinically healthy individuals and ALS samples[31]. Based on histone marks, accessible regions are further classified as a promoter if the region has a histone mark of H3K4me3 and is within 200 base pair (bp) vicinity of the annotated transcript start site, an active enhancer if the region has both H3K27ac and H3K4me1 mark, and a poised enhancer if only H3K4me1 signal is present while H3K27ac mark is lacking (Fig. 2a). The largest category of accessible regions by ChIP-seq marks is active enhancers ($\mathrm{n} = 39{,}559$) with H3K27ac and H3K4me1 followed by thirst largest category of $18{,}817$ poised enhancers regions with only H3K4me1 mark (Fig. 2b). There are $16{,}446$ promoter-like accessible regions with all three H3K4me3, H3K27ac, and H3K4me1 histone marks. Both over-accessibility and under-accessibility outliers are significantly enriched ($\mathrm{P} = 9.2 \times 10^{-3}$ and 7.1×10^{119} respectively based on Fisher's exact test) for promoter-like regions because 28% of the outliers are located in promoter regions while only 14.3% of accessible regions are promoters (Fig. 2b-d). Active enhancers are the second largest category of outliers ($n = 951$) based on the annotation and are not significantly enriched or depleted for outliers ($\mathrm{P} = 0.06$ for under-accessibility and $\mathrm{P} = 0.7$ for over-accessibility based on Fisher's exact test). Furthermore, under-accessibility outliers for both poised enhancer and unannotated regions strongly depleted for outliers as expected ($\mathrm{P} = 1.9 \times 10^{-2}$, 10^{-31} respectively based on Fisher's exact test). Significant enrichment of outliers in functional regions indicates the potential importance of accessibility outliers on the molecular phenotype of ALS disease.

Another interesting observation is that outliers co-occur in the vicinity of each other (Fig. 2e). For example, 7% of outliers have a second outlier in the 10 kilo-bp vicinity with an odds ratio of 688 ($\mathrm{P} = 0$ based on the Fisher's exact test), and 33% of outliers have a second outlier in 1 million bp with the odds ratio of 33 ($\mathrm{P} = 0$ based on the Fisher's exact test). We repeated enrichment analysis between promoter outliers and active or poised enhancer outliers and again observed significant enrichment. 16% of promoter outliers have at least one active enhancer outlier in 100 kilo-bp vicinity ($\mathrm{odds\ ratio} = 403$, $\mathrm{P} = 0$ based on the Fisher's exact test), and 33% of promoter outliers have an

active enhancer outlier in 1 million bp vicinity ($\text{odds ratio} = 74$, $\mathrm{P} = 0$ based on the Fisher's exact test). The co-occurrence of the outliers indicates the potential interaction between them. Hi-C experiments from motor neurons provide further evidence for the potential interaction between outliers. For example, the relatively high Hi-C score between the outlier promoter of the *ZFP41* gene and a distal enhancer outlier located ~ 110 kilo-bp upstream of the gene suggests a potential interaction between outliers (Fig. 2f). We calculated the Hi-C score between pairs of accessible regions and categorized regions by outlier status. We observed that outlier pairs ($\mathrm{P} = 0$ based on the Mann–Whitney U test), including promoter-active enhancer pairs ($P = 4 \times 10^{-44}$ based on the Mann–Whitney U test), have higher interaction scores compared to a baseline where at least one of the regions in the pair is not an outlier (Fig. 2g). Hi-C contact scores are distance-dependent and decay according to power-law with increasing genomic distance. Thus, a higher interaction score could be possibly confounded by the co-occurrence of outliers in the vicinity of each other. To avoid this potential bias, we fit power regression on Hi-C scores using the outlier status feature and distance control variable (Methods). Outlier pairs have higher interaction scores ($P = 0$ based on the t-test) even after controlling for distance with power regression (Extended Data Fig. 5a). Even when we restricted our analysis to region pairs at least 100,000 bp apart, the Hi-C contact scores of outlier pairs still surpass the baseline of non-outlier pairs (Extended Data Fig. 5b). Overall, the co-occurrence of outliers in the vicinity of each other and higher Hi-C contact scores between outlier pairs indicate a potential interaction between outliers.

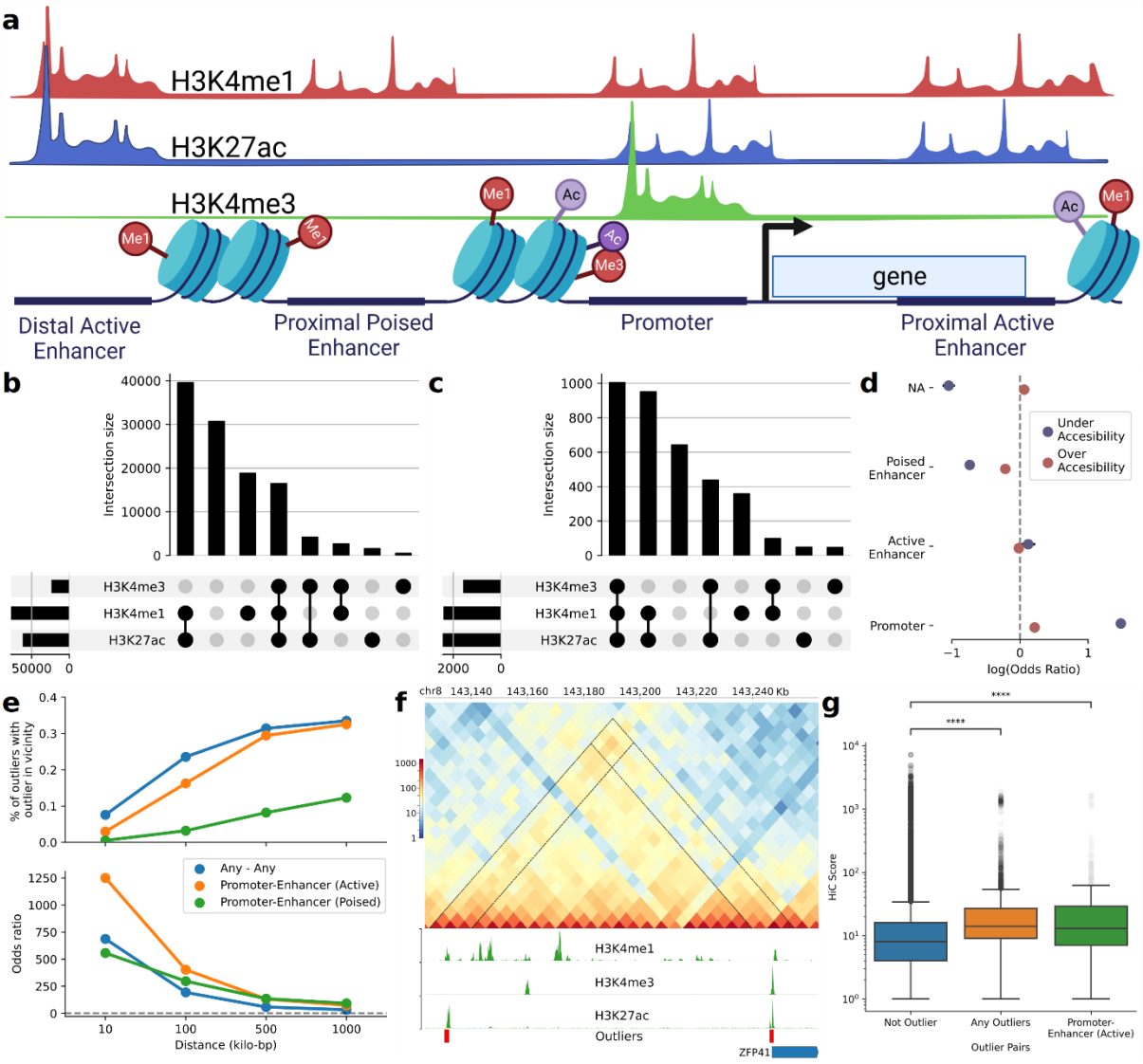

Fig. 2: (a) Accessible regions are annotated using ChIP-seq marks and gene annotation. Then, accessible regions are annotated as promoters, active or poised, and proximal or distal enhancers. **(b)** An upset plot shows the overlap size between H3K4me3, H3K27ac, and H3K4me1 histone marks and accessible regions **(c)** and outliers. **(d)** Enrichment of over-accessibility and under-accessibility outliers in the promoter, poised or active enhancer, and unannotated regions in terms of the log of the odds ratio. **(e)** Cumulative odds of observing the second outlier in a region given that there is an outlier in the region (top) and the cumulative percentage of outliers with a second outlier in the vicinity (bottom) based on the distance between regions and the annotation. **(f)** The interaction between the outlier promoter of ZFP41 and ~100 kilo-bp apart outlier distal enhancer is highlighted by the Hi-C track containing contact score between 5 kilo-bp long genomic bins. Coverage tracks for H3K4me3, H3K27ac, and H3K4me1 histone marks are also included in the figure. Red boxes indicate the outlier status of the accessible regions. **(g)** Boxplot shows the Hi-C contact scores distribution of non-outlier, outlier, and promoter-enhancer pairs. P-values are calculated with the Mann–Whitney U test and corrected with the Bonferroni correction for multiple testing.

DNA accessibility outliers influence gene expression and protein levels

Aberration in accessibility can impact downstream molecular phenotypes such as gene expression and protein levels. For example, the promoter of the *LCMT1* gene is an under-accessibility outlier in an ALS sample (Fig. 3a-c). We investigated a similar aberration pattern in the gene expression and protein level of *LCMT1* from the same sample's RNA-seq and mass-spectrometry experiments. We observed that there is a significant decrease in gene expression ($\text{Z-score} = -11.1$, $\text{P} = 2.8 \times 10^{-20}$ based on the Negative Binomial test). Moreover, the reduction in protein levels ($\text{Z-score} = -5.4$) is potentially linked to the under-accessibility of the promoter (Fig. 3d,e).

To understand the relationship between accessibility outliers and gene expression, we compared the variations in promoter accessibility with variations in gene expression across all samples (Fig. 3f). We observed a significant correlation between the fold changes ($log_2(FC)$) in accessibility at genes' promoters and the respective expression levels of these genes (Spearman's correlation coefficient=66%, $\text{P} = 0$). The high correlation indicates that aberrations in promoter accessibility potentially influence the aberrations in gene expression.

Moreover, we demonstrated that gene expression outliers can be systematically predicted from the promoters, proximal, and distal enhancer accessibility (Fig. 3g). Ranking promoter outliers by their absolute log fold change ($|\log_2(FC)|$) to predict gene expression outliers achieve an auPRC of 14%. If the promoter is an outlier, there is a 41% chance (precision) that its gene is an expression outlier, and 26% of gene expression outliers have an outlier promoter (recall). Similarly, proximal enhancer outliers are highly predictive of gene expression outliers. Specifically, 35% (the recall at 19% precision) of gene expression outliers have at least one outlier proximal enhancer. Ranking genes based on the absolute log fold change of their outlier proximal enhancers achieves the performance of auPRC of 7.8%. We also ranked genes based on the maximum absolute fold change of accessibility outliers in 100 kilo-bp vicinity regardless of annotation of outliers and achieved 3.2% auPRC. We weighted the absolute log fold change of distal enhancers by ABC score[32] and obtained a score for each gene (Methods). The score calculated from distal enhancer outliers is also a reliable predictor of gene expression outliers and achieves 5.9% auPRC. We trained an explainable boosting machine[33] to predict gene expression outliers by combining features from promoters, proximal, and distal enhancers (Methods). The machine learning model achieved 21.4% auPRC. This analysis shows that 59% of the gene expression outliers (with 1% precision) can be linked to aberrations in accessibility because those gene expression outliers have at least one promoter, proximal, or distal enhancer outliers in the vicinity.

We focused on gene expression outliers with outlier promoters and investigated aberrations in their protein levels (Fig. 3h). We observed that the gene expression outliers with overly accessible

promoters have higher protein levels (average Z-score $= 2.4$, $\mathrm{P} = 1.3 \times 10^{-7}$ based on the Mann–Whitney U test), and genes with under-accessible promoters have lower protein levels (average Z-score $= -2.6$, $\mathrm{P} = 2 \times 10^{-11}$ based on the Mann–Whitney U test). Overall, we present the biological significance of accessibility outliers on molecular phenotype by replicating outliers from multiple omics levels.

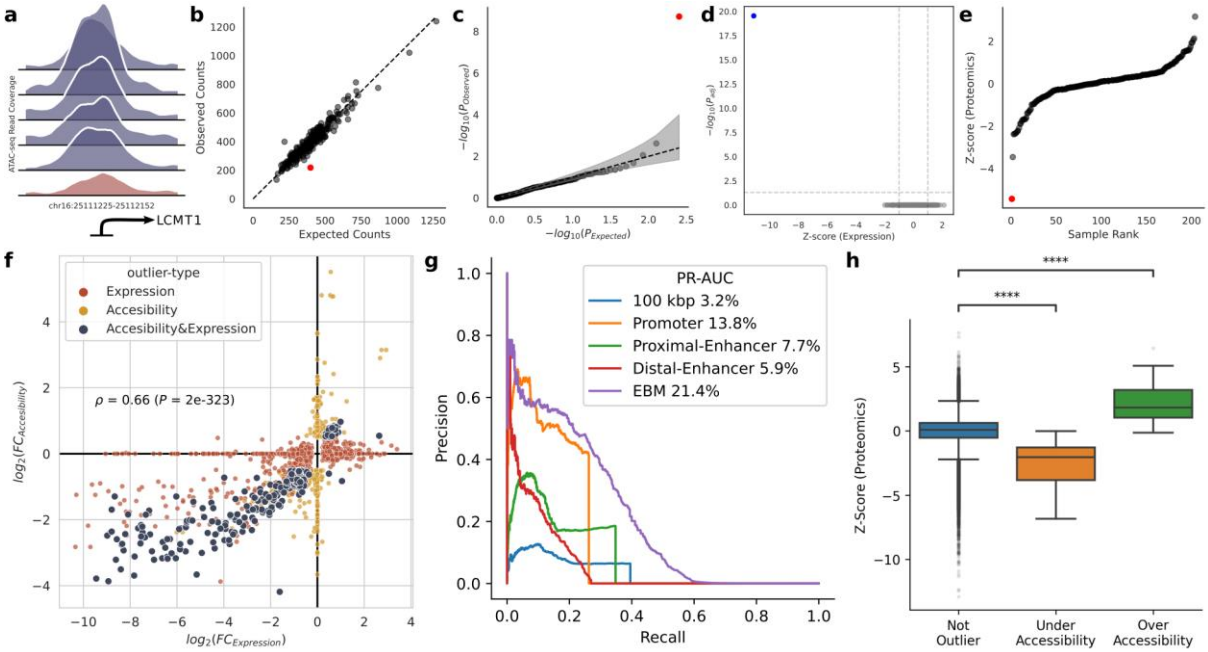

Fig. 3: Prediction of gene expression outliers and aberrant protein levels from DNA accessibility. (a) The outlier promoter of the *LCMT1* gene (in red) has a much lower ATAC-seq read coverage than promoters of non-outlier samples. **(b)** Expected and observed accessibility in the promoter of LCMT1 gene across samples. The outlier sample is indicated with a red dot in the figure panels. **(c)** QQ-plot of expected and observed p-values of the promoter. P-values were calculated using the negative binomial test, and adjusted p-values were obtained using the Benjamini-Yekutieli false-discovery rate (FDR) method. **(d)** The volcano plot shows the aberrant gene expression of *LCMT1* across samples with Z-score and adjusted p-values, which were calculated using a negative binomial test and the Benjamini-Yekutieli false-discovery rate (FDR) method. **(e)** The rank plot of samples by the Z-score displays the aberrant protein levels of *LCMT1* across samples. **(f)** The precision-recall curve reports the performance of a range of predictors to estimate gene expression outliers. Those predictors are the absolute log fold change of promoter outliers (orange), the absolute log fold change of proximal enhancers (green), the maximum absolute fold change of any outlier in 100 kilo-bp vicinity (blue), the absolute log fold change of distal enhancers weighted by ABC score (red), explainable boosting machine (EBM) trained with promoter, proximal-enhancer, and distal-enhancers features (purple). **(g)** The scatterplot shows the correlation between the absolute log fold change of accessibility and expression outliers. **(h)** The boxplot shows the Z-score distribution of proteins, which are categorized by the outlier status of the promoter that transcripts them.

334 **A comparison of DNA accessibility and gene expression reveals if aberrations in molecular**
335 **phenotype are linked to pre-transcriptional or post-transcriptional regulation**

336 The transcriptional synthesis and degradation rate control the gene expression (Fig. 4a). Pre-
337 transcriptional regulation, such as aberrant promoter or enhancer activity, can lead to up or down-
338 regulation of gene expression by altering the synthesis rate. Alternatively, post-transcriptional
339 processes disrupted by genetic variants may lead to the degradation of transcripts via a non-sense-
340 mediated decay. The potential regulatory process associated with aberrant gene expression can be
341 revealed by comparing accessibility outliers against gene expression outliers.

342 To demonstrate this point, we investigated the enrichment of rare high-impact variants in different
343 outlier categories. Fig. 4b presents the frequency of high-impact rare variants, such as splice acceptor,
344 donor, nonsense, and frameshift variants, based on their SnpEff[34] consequences by the outlier type of
345 the affected gene. High-impact variants are observed rarely and appear in fewer than 0.1% of non-
346 outlier genes. Compared to the baseline of all expression outliers, high-impact genetic variants are
347 significantly more prevalent in gene expression outliers if their promoters are not accessibility outliers.
348 The enrichment holds for splicing acceptor $(\mathrm{n} = 11,\ \mathrm{P} = 5.4 \times 10^{-4}$ based on the Mann–Whitney U
349 test calculated with bootstrapping) and splicing donor $(\mathrm{n} = 11, \mathrm{P} = 3.8 \times 10^{-4})$, nonsense $(\mathrm{n} =$
350 $39, \mathrm{P} = 2.5 \times 10^{-7})$, and frameshift variants $(\mathrm{n} = 52, \mathrm{P} = 1.1 \times 10^{-7})$. Remarkably, none of the gene
351 expression outliers with aberrant promoter activity contain high-impact variants, indicating a
352 significant depletion pattern.

353 Further investigation on both moderate-impact missense and high-impact variants on the protein
354 levels presents a similar trend of depletion of variants in genes with aberrant promoter accessibility
355 (Fig. 4c). 28.7% of expression outlier genes $(n = 101)$ with aberrant protein levels $(|\text{Z-score}| > 1)$
356 contain at least one high-impact variant or missense variant. When genes with promoter outliers were
357 excluded, the enrichment of variants rose to 34.1% $(n = 85, P = 0$ based on the Mann–Whitney U
358 test calculated with bootstrapping). The remaining expression outlier genes $(n = 16)$ have promoters
359 with aberrant accessibility, and none of these genes contain genetic variants with high or moderate
360 consequences. The substantial depletion of high-impact variants in these genes $(P = 0)$ indicates that
361 their aberrant protein levels are potentially linked to aberrant promoter accessibility rather than
362 coding variants.

363 Splicing is another mechanism that regulates gene expression[35]. Thus, we further explored the impact
364 of exonic or intronic splicing-disrupting variants prioritized by AbSplice[36]. We detected 43 gene
365 expression outliers containing at least one splicing-disrupting variant across different thresholds (Fig.
366 4d). However, none of these genes has an outlier promoter regardless of the threshold choice. In

contrast, subsetting gene expression outliers without aberrant promoters increases the prevalence of splicing disturbing variants for the subset.

The results underscore that comparing accessibility outliers against gene expression outliers and aberrant protein levels can identify whether aberration in the molecular phenotype is tied to pre-transcriptional or post-transcriptional regulation.

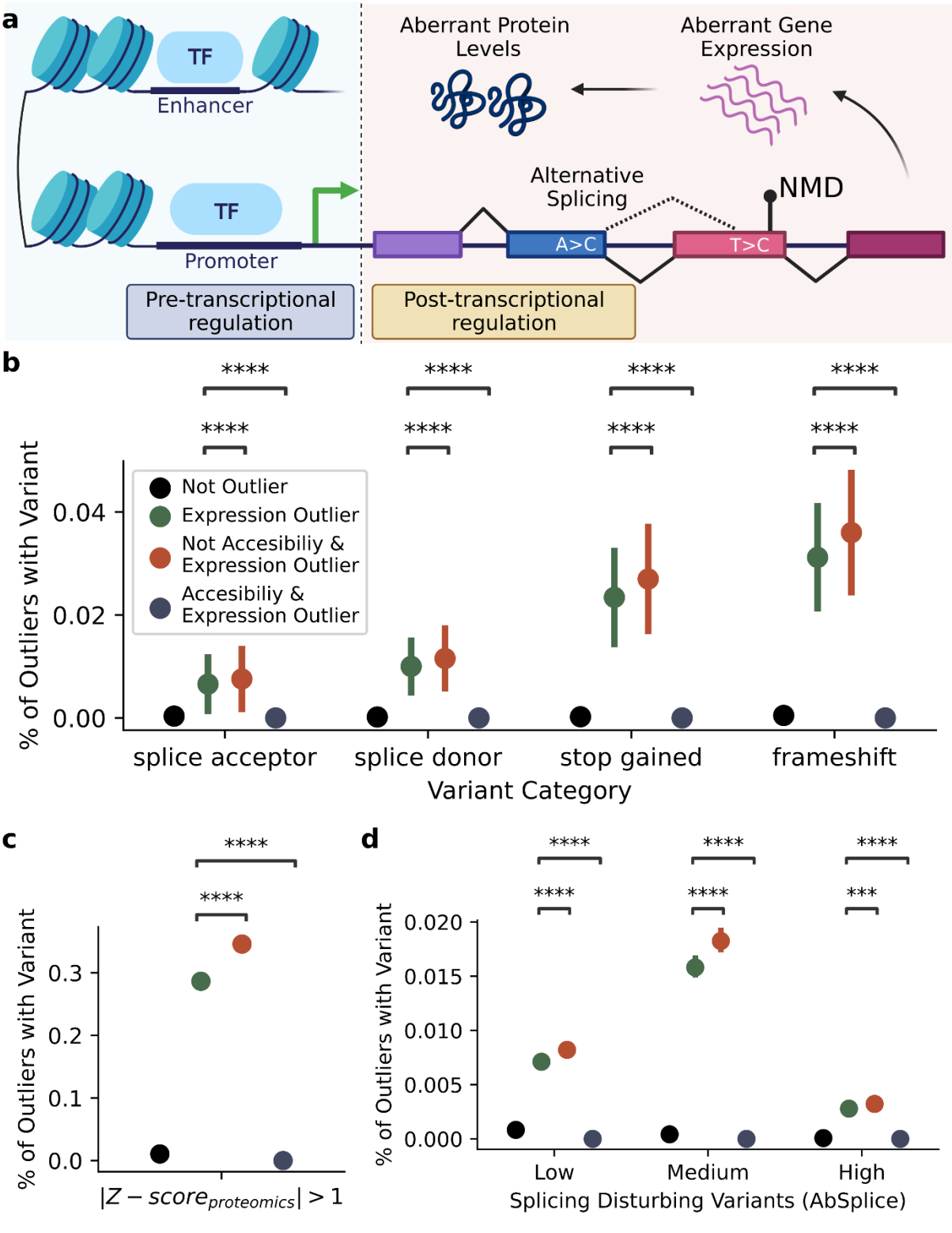

Fig. 4: Overlap between accessibility outliers, gene expression outliers, and high-impact genetic variants. (a) Aberration in gene expression can result from dysregulation in pre-transcriptional regulation, such as aberrant promoter or enhancer activity, or dysregulation in post-transcriptional regulation, such as splicing or nonsense-mediated decay. **(b)** Percentage of genes with high-impact genetic variants by outlier categories. Gene expression outliers without outlier promoters are enriched, and gene expression outliers with promoter outliers are depleted for variants in each category. **(c)** Percentage of protein with a missense or high-impact variant where protein levels are at least deviant by $|Z\text{-score}| > 1$ by outlier categories. **(d)** Percentage of genes containing splicing-disrupting variants predicted by AbSplice by outlier categories.

Discussion

In this study, we introduced a computational toolbox for detecting and annotating DNA accessibility outliers, which are characterized as large aberrations in a few regions specific to a single or few samples. We applied the toolbox to ATAC-seq data from a clinical health and ALS sample cohort. Our methodology employs a negative binomial test for detecting outliers with statistical significance. The mean parameter of the negative binomial is fitted using a linear autoencoder (LR-AE), and the dispersion parameter is inferred based on observed and expected counts using maximum likelihood estimation. Our software, EpiOut, is highly scalable and can handle unknown confounding factors. While primarily designed for ATAC-seq, our toolbox is readily compatible with other accessibility assays, such as DNase-seq.

Our toolbox includes EpiAnnot, which annotates accessible regions for ChIP-seq marks. Using EpiAnnot, we found that outliers are enriched in functional regions, particularly promoters and active enhancers. Interestingly, outlier pairs tend to occur nearby, with many promoter outliers tied to active enhancer outliers within one million bp vicinity. This observation is supported by the relatively high Hi-C scores for outlier pairs in the vicinity, indicating potential interactions between these outliers.

By examining multiple omics levels, we found consistent replication of outliers. Accessibility outliers are associated with downstream biological processes such as gene expression and protein levels. In particular, a significant proportion of the gene expression outliers can be predicted from the aberrant accessibility of the promoter, proximal, and distal enhancer regions. Similarly, aberrant promoter activity is correlated with up and down-regulation of protein levels.

Analyzing the interplay between accessibility and gene expression outliers yields insight into whether aberration in gene expression originates from pre-transcriptional regulation, such as increased synthesis rate via higher promoter activity, or post-transcriptional regulation, such as splicing or nonsense-mediated decay. We observed substantial depletion of high-impact variants in gene

407 expression outliers if promoters of these genes are an accessibility outlier and conversely observed
408 enrichment of high-impact variants if their promoter is not an outlier.

409 Our epigenetic outlier detection methodology offers a novel avenue for studying rare diseases. We
410 have successfully adapted the outlier detection approach for ATAC-seq data and underscored that
411 ATAC-seq is a beneficial complementary assay for rare disease diagnostics. Detected outliers in ALS
412 samples are highly robust and consistently replicated across multiple omics levels. The continued
413 development and integration of the outlier detection approach with disease gene discovery
414 methodologies may ultimately lead to a more comprehensive understanding of the genetic factors
415 contributing to ALS. Such advancements could finally bridge the gap between the disease heritability
416 and the known catalog of ALS disease genes.

417

418 ## Method

419 **AnswerALS Dataset**

420 The multi-omics dataset, which includes ATAC-seq, RNA-seq, proteomics, and whole genome
421 sequencing (WGS), for Amyotrophic Lateral Sclerosis (ALS) was downloaded from the Answer ALS
422 portal (dataportal.answerals.org). The data contains 245 individuals diagnosed with ALS and 45
423 samples from clinically healthy controls. All samples had corresponding ATAC-seq and RNA-seq
424 experiments. 253 samples also have whole genome sequencing (WGS) data. For the scope of our
425 study, we restricted our analysis to the 253 samples that had paired data across three omics levels.
426 Additionally, 204 of these samples have proteomics data. In the proteomics-related analysis, we
427 subset and only used these samples with the proteomics data.

428

429 **Peak Calling**

430 We performed joint peak calling on ATAC-seq data to detect accessible regions across all samples
431 using MACS2[26]. First, we merged bam files from every sample into a unified bam file utilizing
432 SAMtools[37], and subsequently filtered out reads with a mapping quality (MAPQ) below 10. Duplicate
433 reads were retained after merging reads across samples. The default arguments of MACS2 were
434 used, except for the duplicate read filter. ATAC-seq peaks contained in the narrow peak bed file
435 generated by MACS2 were used in the downstream analysis.

436

437 **Read Counting**

438 We implemented an ultra-fast read counting algorithm for ATAC-seq described in Supplementary
439 Algorithm-1. The counting algorithm is optimized based on two primary assumptions: ATAC-seq

peaks are not overlapping and are separated by at least a gap longer than a read length; moreover, both peaks and reads are sorted by the genomic coordinates. We ensure the first assumption by jointly calling peaks as described above and collapsing any peaks closer to each other than the minimum gap distance (default 200 bp); thus, an ATAC-seq read never intersects with two peaks simultaneously. Since the bam file format is pre-sorted by genomic coordinates, ATAC-seq peaks are sorted to guarantee the second condition before the counting step. The counting algorithm creates two stacks of sorted peaks and reads, iterates over reads and peaks, and tracks the number of overlaps (Supplementary Algorithm-1). The resulting runtime complexity of the counting step is $\mathrm{O}(r+p)$ where r is the number of reads, and p is the number of peaks. In the preprocessing step, peaks are sorted, so the overall complexity is $\mathrm{O}(p\ log(p)+r)$. However, since the number of reads is much larger than the number of peaks ($\mathrm{r} > \mathrm{p}$), the algorithm practically behaves in linear runtime in terms of the number of reads. The memory complexity of counting is linear in terms of the number of peaks ($\mathrm{O}(p)$) because reads are fetched iteratively per chromosome due to the indexable file format of BAM.

Filtering Peaks by Replication Rate

We applied two filters to the ATAC-seq peaks detected by MACS2 to ensure consistent peak replication across samples. First, we eliminated peaks with low coverage, specifically those with fewer than 100 reads across all samples. As a subsequent criterion, any peak must be supported by a minimum of 2 reads in at least 50% of all samples.

Size Factor Normalization

To account for coverage differences across the ATAC-seq experiments, we performed size factor normalization on the read counts, a method initially proposed by DESeq2. The size factor s_i for a sample i is defined as:

$$k_i^g = \left(\prod_{j=1}^{n} k_{ij}\right)^{1/n} \tag{1.1}$$

$$s_j = \mathrm{median}\frac{k_{ij}}{k_i^g} \tag{1.2}$$

where k_{ij} represents the number of reads mapped to region j in sample i and k_i^g is the geometric mean of the reads across regions for sample i. The median of k_{ij} to k_i^g ratios is defined as the size factor s_i of sample i.

Outlier Detection

After normalizing the read counts of peaks with size factor normalization, we log-transformed the normalized counts and centered them around zero by subtracting the mean normalized counts:

$$x_{ij} = n_{ij} - \bar{n}_{\mathrm{j}} \tag{2.1}$$

$$n_{ij} = \log\left(\frac{k_{ij} + 1}{s_j}\right) \tag{2.2}$$

Autoencoder is applied on x_{ij} to calculate the expectation of normalized read counts ($\hat{x}_{ij}$). f_e is the encoder function of the autoencoder, which takes observed normalized counts (x_{ij}) as input and calculates major covariates. Major covariates are concatenated with known confounders to obtain latent representation (h), which is decoded back to expected normalized counts ($\bar{x}_{ij}$) with the decoder function f_d:

$$\mathrm{h} = \left[f_e(x_{ij}) | c_{known}\right] \tag{3.1}$$

$$\hat{x}_{ij} = f_d(h) \tag{3}$$

We used linear encoder and decoder functions. The encoder model uses a principal component analysis (PCA), where the encoder weights (W_e) are the rotation matrix of PCA. The decoder weights (W_d) are initialized with linear regression and further trained with the negative binomial objective:

$$f_e(x) = W_e \mathrm{x} \tag{4.1}$$

$$f_d(h) = W_d \mathrm{h} \tag{4.2}$$

Log normalized counts (x_{ij}) is transformed back to the original natural scale:

$$\hat{k}_{ij} = e^{\hat{x}_{ij} + n_j} s_j - 1 \tag{5}$$

The dispersion parameter is optimized with maximum likelihood estimation (MLE) where the likelihood function:

$$\mathrm{L}(r_j) = \prod_i NB\left(k_{ij} | \hat{k}_{ij}, r_j\right) \tag{6}$$

Fitting the dispersion parameter requires solving an independent convex problem for each peak. To solve a large number of independent convex problems quickly, we implemented a vectorized backtracking line search algorithm using TensorFlow. The dispersion parameter range is set between a lower bound of 0.01 and an upper bound of 1,000 to avoid numerical stability issues and overfitting.

P-value (P_{ij}) of each peak and sample is calculated with a two-sided negative binomial test:

$$P_{\mathrm{ij}} = \min\left\{\frac{1}{2}, \sum_{\mathrm{k}=0}^{\hat{\mathrm{k}}_{\mathrm{ij}}} \mathrm{NB}\left(\mathrm{k}_{\mathrm{ij}} | \hat{\mathrm{k}}_{\mathrm{ij}}, \mathrm{r}_{\mathrm{j}}\right), 1 - \sum_{\mathrm{k}=0}^{\hat{\mathrm{k}}_{\mathrm{ij}}-1} \mathrm{NB}\left(\mathrm{k}_{\mathrm{ij}} | \hat{\mathrm{k}}_{\mathrm{ij}}, \mathrm{r}_{\mathrm{j}}\right)\right\} \tag{7}$$

Finally, p-values are corrected for multiple testing corrections to control the false discovery rate with the Benjamini-Yekutieli method.

In addition to p-values, we report the log of fold-change or the log ratio of observed read counts to expected read counts calculated by:

$$l_{ij} = \log_2(FC) = \log_2\left(\frac{k_{ij}+1}{\hat{k}_{ij}+1}\right) \tag{8}$$

and Z-score based on fold changes defined as:

$$Z\text{-score}_{ij} = \frac{l_{ij} - \mu_j^l}{\sigma_j^l} \tag{9}$$

We consider peaks as outliers if their adjusted p-values are smaller than 5% ($P_{adj} < 0.05$), their absolute log fold changes are greater than 50% ($|\log_2(\mathrm{FC})| > 0.5$) and either their read counts (k_{ij}) or expected read counts ($\hat{k}_{ij}$) are at least 50.

Injection of artificial outliers

In order to benchmark the performance of outlier detection methods, we conducted a simulation experiment using artificial outliers. We created an outlier mask M_{ij} for sample-peak pairs. A sample-region pair is categorized as either an over-accessibility or under-accessibility outlier with a probability of 0.01%:

$$\mathrm{f}(x) = \begin{cases} 1, & p = 5\ 10^{-4} \\ -1, & p = 5\ 10^{-4} \\ 0, & otherwise \end{cases} \tag{10}$$

Log-normalized counts are updated using an artificial outlier mask:

$$x_{ij}^{inj} = x_{ij} + \sigma_j e^{\mathcal{N}(3,1)} M_{ij} \tag{11}$$

where the deviation in accessibility outlier is simulated by scaling the standard deviation of the peak (σ_j) using a value sampled from a log-normal distribution parameterized by a mean of 3 and a standard deviation of 1. Eq. 5 transforms log-transformed normalized injected counts (x_{ij}^{inj}) back to counts in the natural count scale (k_{ij}^{inj}).

Outlier detection methods are evaluated by benchmarking their capability to predict the outlier mask (M_{ij}) from the injected counts (x_{ij}^{inj}). The primary benchmark metric for this evaluation is the area under the precision-recall curve (auPRC). We generated ten outlier masks for testing and one

523 additional mask for the validation set. Hyperparameter tuning was performed on the validation set
524 to identify the optimal bottleneck size of autoencoder-based models. The evaluation procedure was
525 executed ten times on the test folds, and the average auPRC performance and its standard deviation
526 are reported.

527 We evaluated four methods: naive negative-binomial test, PCA, OUTRIDER, and EpiOut. During the
528 precision-recall calculation, predictions for each method were primarily ranked using p-values,
529 except for PCA, which was ranked by its Z-score. The performance of each method at a specific
530 cutoff (either at the adjusted p-values of 0.05 or the absolute Z-score of 2) is indicated on the
531 precision-recall curve.

532 For the naive negative binomial test, we averaged counts across samples per peak to obtain the
533 mean parameter of the negative binomial distribution, $NB\left(k_{ij}^{inj} | \bar{k}_{ij}^{inj}, r_j\right)$. The dispersion is
534 estimated with MLE (Eq. 6).

535 To evaluate the PCA model, we used PCA to estimate the expected read counts, $\hat{k}_j^{inj}$. These
536 expected read counts are the reconstructed read counts of PCA. The number of principal
537 components retained was determined through hyperparameter tuning on the validation set.
538 Predictions were subsequently ranked based on the Z-score, as formulated in Eq. 9, using the
539 expected counts from PCA. We did not compute dispersion or perform a negative binomial test.

540 We ran OUTRIDER and EpiOut with their default parameters. The bottleneck size of all autoencoder-
541 based methods was selected based on the hyperparameter tuning on the validation set.

542 **Runtime Benchmark**

543 To measure runtime to call outliers with OUTRIDER and EpiOut, we subset peaks into groups of
544 10000, 25000, 50000, and 100000 and calculated runtime for each input size. 8 CPU cores are used
545 during the benchmark. The analysis is repeated 10 times. The average runtime and standard
546 deviation of runtime were reported. We ran each tool with bottleneck sizes of 5, 10, 50, and 100 in
547 each iteration to avoid runtime differences resulting from hyperparameter tuning. We reported the
548 total runtime across different bottleneck sizes.

549 **Functional annotation of accessible regions with ChIP-seq**

550 We introduced EpiAnnot, a software tool for functional annotation of accessible genomic regions
551 using ChIP-seq marks. If a specific ChIP-seq for a cell line or tissue is present in the Roadmap

552 Epigenomics[29] or ENCODE[30] databases, EpiAnnot can retrieve ChIP-seq data from these public data
553 sources. Additionally, users can annotate their accessible regions using their own custom ChIP-seq
554 data.

555 EpiAnnot can also attribute accessible regions to genomic features using a gene transfer format
556 (GTF) file. For instance, it annotates accessible regions as 5' UTR, TSS, etc. In scenarios where an
557 accessible region overlaps with the H3K4me3 histone ChIP-seq mark and TSS of a gene, EpiAnnot
558 designates these regions as promoters. Conversely, regions intersecting with the H3K4me1 histone
559 ChIP-seq mark are labeled as enhancers. When the histone mark is available, enhancers overlapping
560 with H3K27ac are categorized as active; otherwise, they are poised. EpiAnnot also annotates
561 enhancers as proximal or distal based on their distance from genes. Specifically, enhancers within a
562 gene body or located up to 10,000 bp upstream or 2,000 bp downstream from a gene are tagged as
563 proximal. All others are marked as distal enhancers.

564 By using EpiAnnot with H3K4me3, H3K27ac, and H3K4me1 histone ChIP-seq marks from in vitro
565 differentiated motor neurons from clinically healthy and ALS samples[31], we annotated accessible
566 regions from ATAC-seq experiments and used the GTF file of GENCODE v38. The histone ChIP-seq
567 data was downloaded from ENCODE.

568 **Enrichment of Outlier Pairs in the Proximity**

569 To investigate the potential interaction between nearby outlier pairs, we quantified the pairs of
570 accessible regions within predefined distances, irrespective of the outlier status of regions. Then, we
571 applied Fisher's exact test to calculate the enrichment of outlier pairs, using a contingency table
572 structured with the outlier statuses of accessible regions in the pair. We repeated this analysis for
573 10,000, 100,000, 500,000, and 1,000,000 bp distances. The odds ratio was also calculated to
574 highlight the likelihood of the region being an outlier, given the outlier status of the nearby region.

575 **Annotation of Interactions with Hi-C**

576 To delve deeper into the interactions between outliers, we annotated their interactions based on Hi-
577 C contacts using EpiAnnot. We downloaded public Hi-C data from motor neurons differentiated from
578 iPSCs[31] from ENCODE and computed interaction scores between accessible regions using EpiAnnot.
579 For the interaction analysis, the genome is partitioned into bins of 5,000 bp. The interaction
580 between a first source region (s) and a target second region (t) is determined by the highest Hi-C
581 score between these bins or their immediate neighbors:

$$h_{s\to t} = \max\{HiC_{score}(bin_i, bin_j), i \in \{t-1, t, t+1\}, j \in \{s-1, s, s+1\}\} \quad (12)$$

We utilized EpiAnnot to compute the ABC score, which signifies potential interactions between regions informed by Hi-C scores. The ABC score is defined as:

$$\mathrm{ABC}_{s\to t} = \frac{h_{s\to t} * k_t}{\sum_{i \in vicinity(s)} h_{s\to i} * k_i} \quad (13)$$

where the numerator represents the Hi-C score between regions multiplied by the accessibility of the target region ($k_t = \sum_i k_{ij}$) and the denominator normalizes this value for all region pairs associated with the source region.

Aberrant Gene Expression Prediction from DNA Accessibility

Gene expression outliers are called with OUTRIDER using the DROP pipeline. Those gene expression outliers constitute the ground truth labels. In order to predict gene expression outliers from accessibility outliers, we trained an explainable boosting machine (EBM). The EBM model was trained and tested with 5-fold cross-validation. The model aggregates features from promoters and enhancers as input to predict the probability of a gene being categorized as an expression outlier. For predicting a gene's outlier status, the model considers the following features: log fold change, p-value, and outlier status of the promoter, maximum absolute log fold change of the proximal enhancer outliers, and the ABC-score weighted absolute log fold change of the distal enhancer outliers. The score for the weighted distal enhancer is computed as follows:

$$\mathrm{enhancer}_{distal} = \sum_{enhancer \in g} \left| ABC_{promoter\to enhancer} * log_2(FC_{enhancer}) \right| \quad (14)$$

Depletion of high-impact variants in Genes with Promoter Outlier

We assessed the depletion of high-impact variants in the gene expression outliers where promoters of these genes are also outliers. Variants are annotated with SnpEff v4.3[34] according to their consequences for genes. Only rare variants with a frequency of 0.1% are in the analysis. Variant frequencies are downloaded from the gnomAD database[38]. Consequence categories used for depletion analysis include stop codon, frameshift, acceptor, donor site dinucleotides, and missense variants. AbSplice was used to classify splicing-disrupting exonic or intronic variants into high, medium, or low-impact categories, using thresholds of 0.2, 0.1, and 0.05, respectively, as advised by the authors. We determined the proportion of genes affected by each variant category based on the outlier status of the genes and their associated promoters. We repeated the enrichment analysis 100 times by randomly subsetting genes in each iteration for bootstrapping. We used the Mann–Whitney U test to measure statistical significance based on the proportions calculated with

612 bootstrapping. Moreover, we analyzed proteomics data to quantify the proportion of proteins
613 impacted by high-impact or missense variants. Z-scores for the proteomics data were obtained with
614 PROTRIDER.

615 **Source Code**

616 All the analyses in the paper are implemented in the reproducible snakemake[39] format available at
617 **github.com/uci-cbcl/ALS-accessibility-outliers-paper**. Epiout, the Python package for outlier
618 detection and annotation, is available at **github.com/uci-cbcl/EpiOut**.

619 **Acknowledgments**

620 This project is funded by a grant from the Robert Packard Center for ALS Research at Johns Hopkins
621 and grants from the National Science Foundation [IIS-1715017, DMS-1763272].

622 ## References

623 1. Hardiman, O. *et al.* Amyotrophic lateral sclerosis. *Nat. Rev. Dis. Primer* **3**, 1–19 (2017).
624 2. Longinetti, E. & Fang, F. Epidemiology of amyotrophic lateral sclerosis: an update of recent
625 literature. *Curr. Opin. Neurol.* **32**, 771–776 (2019).
626 3. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur. J. Neurol.* **27**,
627 1918–1929 (2020).
628 4. Al-Chalabi, A. *et al.* An estimate of amyotrophic lateral sclerosis heritability using twin data. *J.*
629 *Neurol. Neurosurg. Psychiatry* **81**, 1324–1326 (2010).
630 5. Grassano, M. *et al.* Systematic evaluation of genetic mutations in ALS: a population-based study.
631 *J. Neurol. Neurosurg. Psychiatry* **93**, 1190–1193 (2022).
632 6. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat.*
633 *Neurosci.* **17**, 17–23 (2014).
634 7. Van Rheenen, W. *et al.* Project MinE: study design and pilot analyses of a large-scale whole-
635 genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* **26**, 1537–1546
636 (2018).
637 8. Baxi, E. G. *et al.* Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical
638 and multi-omics data from induced pluripotent cell lines. *Nat. Neurosci.* **25**, 226–237 (2022).
639 9. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral
640 sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat.*
641 *Genet.* **53**, 1636–1648 (2021).
642 10. Gregory, J. M., Fagegaltier, D., Phatnani, H. & Harms, M. B. Genetics of Amyotrophic Lateral
643 Sclerosis. *Curr. Genet. Med. Rep.* **8**, 121–131 (2020).
644 11. NeuroLINCS Consortium *et al.* An integrated multi-omic analysis of iPSC-derived motor
645 neurons from C9ORF72 ALS patients. *iScience* **24**, 103221 (2021).
646 12. Frésard, L. *et al.* Identification of rare-disease genes using blood transcriptome sequencing
647 and large control cohorts. *Nat. Med.* **25**, 911–919 (2019).
648 13. Mertes, C. *et al.* Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat.*
649 *Commun.* **12**, 529 (2021).

14. Yépez, V. A. *et al.* Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* **16**, 1276–1296 (2021).
15. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
16. Marwaha, S., Knowles, J. W. & Ashley, E. A. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med.* **14**, 23 (2022).
17. Jenkinson, G. *et al.* LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics* **36**, 4609–4615 (2020).
18. Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
19. Kopajtich, R. *et al.* Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. 2021.03.09.21253187 Preprint at https://doi.org/10.1101/2021.03.09.21253187 (2021).
20. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).
21. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol.* **23**, 210 (2022).
22. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
23. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20**, 295 (2019).
24. Brechtmann, F. *et al.* OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* **103**, 907–917 (2018).
25. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.
26. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
27. Stovner, E. B. & Sætrom, P. PyRanges: efficient comparison of genomic intervals in Python. *Bioinforma. Oxf. Engl.* **36**, 918–919 (2020).
28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
29. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
30. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
31. Zhang, S. *et al.* Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron* **110**, 992-1008.e11 (2022).
32. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
33. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. Preprint at https://doi.org/10.48550/arXiv.1909.09223 (2019).
34. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
35. Daguenet, E., Dujardin, G. & Valcárcel, J. The pathogenicity of splicing defects: mechanistic insights into pre-mRNA processing inform novel therapeutic approaches. *EMBO Rep.* **16**, 1640–1655 (2015).
36. Çelik, M. H. *et al.* Aberrant splicing prediction across human tissues. Preprint at https://doi.org/10.1101/2022.06.13.495326 (2022).

37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
39. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.* **28**, 2520–2522 (2012).