Unveiling causal regulatory mechanisms through cell-state parallax

Alexander Po-Yen Wu[†1], Rohit Singh[†*1],Christopher Walsh[2], Bonnie Berger[*1,3]

[†] Contributed equally.

* Correspondence: rsingh@mit.edu, bab@mit.edu

Abstract

Genome-wide association studies (GWAS) identify numerous disease-linked genetic variants at noncoding genomic loci, yet therapeutic progress is hampered by the challenge of deciphering the regulatory roles of these loci in tissue-specific contexts. Single-cell multimodal assays that simultaneously profile chromatin accessibility and gene expression could predict tissue-specific causal links between noncoding loci and the genes they affect. However, current computational strategies either neglect the causal relationship between chromatin accessibility and transcription or lack variant-level precision, aggregating data across genomic ranges due to data sparsity. To address this, we introduce GrID-Net, a graph neural network approach that generalizes Granger causal inference to detect new causal locus–gene associations in graph-structured systems such as single-cell trajectories. Inspired by the principles of optical parallax, which reveals object depth from static snapshots, we hypothesized that causal mechanisms could be inferred from static single-cell snapshots by exploiting the time lag between epigenetic and transcriptional cell states, a concept we term "cell-state parallax." Applying GrID-Net to schizophrenia (SCZ) genetic variants, we increased variant coverage by 36% and uncovered noncoding mechanisms that dysregulate 132 genes, including key potassium transporters such as KCNG2 and SLC12A6. Furthermore, we discovered evidence for the prominent role of neural transcription-factor binding disruptions in SCZ etiology. Our work not only provides a strategy

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA
[2] Division of Genetics and Genomics and Howard Hughes Medical Institute, Boston Children's Hospital and Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA, 02115, USA
[3] Department of Mathematics, MIT, Cambridge, MA 02139, USA

for elucidating the tissue-specific impact of noncoding variants but also underscores the breakthrough potential of cell-state parallax in single-cell multiomics for discovering tissue-specific gene regulatory mechanisms.

Introduction

Dynamic changes in the chromatin states of cis-regulatory genomic elements influence gene expression in many biological processes, including development, differentiation, reprogramming, and perturbation response[1–3]. Correspondingly, genome-wide association studies (GWAS) have implicated thousands of noncoding loci in diseases and complex traits[4,5]. Yet, the challenge of linking these noncoding loci to the genes they regulate has impeded therapeutic advances. The common approach of simply associating a locus to its nearest gene is inadequate, as it overlooks distal locus-gene interactions. This is a critical limitation, as distal interactions are likely to be important for many genes. Meanwhile, experimental approaches to discover such interactions can be difficult to scale. Expression quantitative trait locus (eQTL) analyses require sample collections from large cohorts[6], and CRISPR perturbations have been limited to specific loci[7]. Computational methods based on bulk data have also been proposed, such as Wang et al.'s integration of multiple data types to link variants to genes[8], and Fulco et al.'s activity-by-contact (ABC) model[9,10] that predicts enhancers of a gene based on the activity and 3D chromatin structure around the gene. However, these models focus on the relative spatial positions of enhancers and genes, and are not designed to evaluate whether changes in accessibility at a locus actually influence gene expression.

Recent advances in single-cell multimodal technology, which can simultaneously profile chromatin accessibility (scATAC-seq) and gene expression (scRNA-seq) in the same cell, offer a promising opportunity to uncover tissue-specific, genome-scale, causal relationships between noncoding loci and genes in a single experiment[11–14]. However, current computational methods face substantial hurdles in extracting meaningful locus-gene associations. Existing approaches have diverse and often unrelated objectives: achieving concordance between the modalities[15,16], inferring gene regulatory networks by considering the accessibility of transcription-factor binding motifs[17,18], or using chromatin accessibility to improve RNA velocity inference[19]. Unfortunately,

27 owing to the extreme sparsity of scATAC-seq data (< 1% non-zero entries per cell [20]), these
28 approaches aggregate chromatin peaks within a defined range of a gene; for instance, MultiVelo
29 consolidates chromatin signals into a single, per-gene score[19]. This results in a loss of variant-
30 and locus-level resolution and also risks ignoring distal peaks. Moreover, these methods are
31 largely reliant on correlational analyses between scATAC-seq and scRNA-seq, thereby
32 overlooking the inherent causal time lag in gene regulation (**Figure 1a**). For instance, poised
33 enhancers, which play essential roles in development, undergo changes in chromatin states and
34 histone modifications *prior* to activating their target genes and would be missed by a
35 correlational analysis[21]. Consequently, the appropriate epigenetic and transcriptional readouts
36 to be cross-referenced may correspond to subtly different cell states.
37 We refer to this temporal discrepancy as “cell-state parallax,” drawing inspiration from
38 the optical concept of parallax, which describes how different vantage points can unveil new
39 insights. Cell phone cameras with dual lenses leverage optical parallax by capturing multiple 2D
40 images of the same object from slightly offset positions to enable 3D depth perception.
41 Analogously, in single-cell multimodal assays, each modality captures a snapshot of the cell's
42 state, but at an offset due to the temporal lag between the epigenome and the transcriptome. By
43 exploiting this offset, causal relationships between the modalities can be unveiled. However, a
44 fundamental challenge in accounting for these offsets is that single-cell experiments destroy
45 cells upon measurement, so the temporal ordering of single-cell measurements is often unclear.
46 Moreover, an inference framework needs to be robust to the sparsity of both scATAC-seq and
47 scRNA-seq data[22,23]. Some previous methods mitigate sparsity by averaging transcriptionally
48 similar cells into pseudocells[11], but this comes at the cost of losing cell-level granularity which is
49 crucial for discerning subtle offsets between distinct modalities. In this work, we operate at a
50 per-peak and per-cell resolution, addressing sparsity through our machine learning framework.
51 To resolve single-cell multimodal snapshots into causal regulatory relationships, we
52 leverage mathematical principles that underlie the dynamics between causes and their effects.

We focus on Granger causality, a statistical framework which exploits the idea that a cause precedes, and is therefore predictive of, its downstream effect. There, a temporally causal relationship between two time-varying processes can be inferred solely by autoregressive analysis of time-resolved observational data. Granger causality has been a powerful tool in biology, econometrics, and other domains for causal discovery without the need to perform interventions.

However, the classical framework of Granger causality is limited to settings where observations can be ordered along a single time-axis. Mathematically, an ordering of observations can be categorized as *total* or *partial*: in the former, any pair of observations can be ordered unambiguously while for the latter only a subset of observations can be so ordered. Partial orderings characterize many dynamical systems, including disinformation cascades in social media, financial transaction networks, and parallel computing systems. In single-cell biology, when a differentiation trajectory has multiple branches, cell states across distinct branches are not directly comparable, necessitating a partial ordering of cells. Classical Granger causal analysis, being limited to total orderings, therefore needs to be adapted for inferring causal single-cell dynamics.

Our key mathematical advance is a generalization of Granger causality to partial orderings. We exploit the equivalence between such orderings and directed acyclic graphs (DAGs) and design a graph neural network framework to infer causality from the dynamics. We complement this with the conceptual advance of cell-state parallax from which we infer a DAG-based representation of single-cell multimodal dynamics. We combine these innovations into GrID-Net (Granger Inference on DAGs), a novel framework for Granger causal inference on DAGs that we apply to single-cell multimodal data for inferring causal noncoding locus-gene relationships (**Figure 1b,c**). Leveraging lagged message-passing on a graph neural network, GrID-Net resolves the cell-state parallax between the epigenome and transcriptome. We demonstrate that it outperforms existing state-of-the-art approaches for accurately identifying

79 noncoding locus-gene links across a variety of cell types and single-cell multimodal assays. We
80 then use GrID-Net to link fine-mapped genetic variants associated with schizophrenia (SCZ) to
81 the genes they affect. We uncover noncoding mechanisms of gene dysregulation that
82 prominently feature disruptions of neural transcription-factor binding and synaptic pathways as
83 putative drivers of SCZ disease pathogenesis (**Figure 1d**). Our findings demonstrate the power
84 of GrID-Net in uncovering tissue-specific causal mechanisms of disease and development,
85 using data generated from just one sample. Our work also illustrates the utility of multimodal
86 single-cell studies for investigating the impact of genetic variants on function and phenotype.
87 The code for GrID-Net can be found at https://github.com/alexw16/gridnet and the full list of
88 SCZ variant-gene links can be found in **Supplemental Table 1**.
89
90 **Overview of GrID-Net**
91 In general, inferring true causality can be a challenge as it requires perturbations to confirm
92 cause-and-effect relationships and validations to rule out potential confounders. Granger
93 causality is a powerful statistical concept that can *approximate* causality in time-varying
94 processes solely from observational data. It leans on the principle that a cause precedes its
95 effect[24,25]. Given two time-dependent variables x (a potential cause) and y (the effect), if the
96 forecast for y can be improved by accounting for the past values of x, then x is said to Granger
97 cause y since changes in x temporally precede those in y.
98 One of our key contributions is to broaden the set of domains to which Granger causal
99 analyses can be applied. Specifically, we generalize Granger causal inference to dynamical
100 systems that are characterized by partial orderings. To model the asynchrony between cause
101 and effect, we modify standard graph neural networks to allow for lagged message passing.
102 Each node (i.e., a cell state) in the directed acyclic graph accumulates information from its
103 parent nodes (past states), leveraging this lagged information to forecast the present state in the
104 context of the graph-structured dynamics. GrID-Net consists of two forecasting models: a

105 *reduced* model forecasts an effect y using only the past of y and a *full* model forecasts the
106 effect y using both the past of y and the past of the candidate causal variable x. The strength of
107 the Granger causal link between x and y can be evaluated by the statistical F-test which
108 quantifies the improvement in forecasting accuracy of the *full* model over that of the *reduced*
109 model.

110 GrID-Net enables the detection of asynchronous interactions between gene regulatory
111 elements along single-cell trajectories. For each single-cell multimodal dataset, a DAG of cell
112 states is constructed from a k-nearest neighbor graph of cells, with edges oriented in
113 accordance with pseudotime computed after a multimodal integration with Schema[20]. RNA-
114 velocity based edge orientations that capture local dynamics can also be supplied, so long as
115 any resulting cycles are broken. We evaluated the Granger causal strength between each gene
116 and chromatin accessible element ("peak") within one megabase (Mb) of the gene's
117 transcription start site, ranking the peak-gene pairs by their Granger causal strengths
118 (**Methods**).
119

120 **GrID-Net more accurately identifies noncoding locus-gene links**

121 We evaluated GrID-Net on four single-cell multimodal datasets (scATAC-seq and scRNA-seq)
122 that characterize a range of biological processes, including cell differentiation and drug
123 treatment response[11,12,26,27]. For each dataset, we sought to validate against expression
124 quantitative trait (eQTL) loci profiled in matching tissue (**Methods**). As eQTLs represent
125 associations between genetic variation at a locus and the corresponding changes in a gene's
126 expression, they serve as useful genome-wide proxies for perturbational validation of causal
127 interactions between genomic loci and genes.

128 For each gene, we intersected all ATAC-seq peaks within 1 Mb of its transcription start
129 site (TSS) with the set of eQTL variants linked to it, associating each peak with the most
130 significant variant overlapping it. We then evaluated the peak–gene pairs prioritized by GrID-Net

131 and two correlation-based methods used in previous work[11,27], Pearson correlation and
132 pseudocell correlation. The latter computes correlations on measurements aggregated across
133 multiple cells. We also evaluated the ABC model's predictions on bulk data from matching
134 tissues, comparing it against the predictions from the single-cell datasets (**Methods**).
135 We evaluated the accuracy of each method in retrieving the peak–gene pairs that
136 overlap significant eQTLs ($P < 10^{-10}$); our findings were robust to the choice of the eQTL
137 significance threshold (**Figure S1**). GrID-Net was significantly more accurate than the other
138 methods across all datasets (**Figure 2a**), exhibiting 7-27% higher area under the receiver
139 operating characteristic curve (AUROC) in 3 of the 4 datasets. GrID-Net demonstrated the
140 smallest improvement in AUROC for the sci-CAR dataset, although all methods performed
141 poorly on this dataset, likely due to the lower data quality of sci-CAR as the earliest single-cell
142 multimodal assay developed. Meanwhile, the correlation-based methods exhibited near-random
143 performance across all datasets, consistently underperforming not only GrID-Net but also the
144 bulk data-based ABC model. The low accuracy of predictions by correlation-based methods,
145 contrasted with the substantial improvement offered by GrID-Net, demonstrates the potential to
146 unravel much richer regulatory insights from these multimodal assays than has been possible.
147 We next assessed the robustness of the single-cell methods to sparsity in single-cell
148 data. GrID-Net better handled sparse data (**Figure S2**), yielding higher accuracies than the
149 correlation-based methods for peak–gene predictions involving genes with sparse data (non-
150 zero transcripts in fewer than 10% of cells) or sparse peaks (under 2% of cells). Notably,
151 pseudocell correlation failed to consistently outperform standard Pearson correlation on these
152 sparse peak–gene cases, even though the former is explicitly designed to compensate for the
153 sparsity of measurements at the level of individual cells. This may be because pseudocell
154 averaging reduces the effective number of independent observations and thus relinquishes the
155 statistical advantages of single-cell heterogeneity. Altogether, these analyses indicate not only

156 GrID-Net's robustness to the current sparsity of single-cell data but also the potential for its
157 increased effectiveness as the sensitivity of these multimodal assays continues to improve.
158 To further benchmark GrID-Net using functional indicators of noncoding locus–gene
159 interactions, we examined three-dimensional chromatin contacts (from bulk Hi-C data) between
160 gene promoters and noncoding loci. Compared to the two correlation-based methods, we found
161 that GrID-Net more accurately prioritized peak–gene pairs that were in frequent three-
162 dimensional contact (**Figure 2b)**. The ABC method uses bulk Hi-C data itself as an input so we
163 did not evaluate its agreement with chromatin contacts. In future, the GrID-Net framework could
164 be expanded to also utilize Hi-C data when available— as we describe later, incorporating one-
165 dimensional sequence proximity as a signal into GrID-Net improves its results. However, Hi-C
166 data is currently available only for a limited set of tissues[28] and the ability to make regulatory
167 inferences from just one multimodal assay enables the convenient study of rare cell types,
168 cancer mutations and drug responses etc.
169
170 **GrID-Net reveals how chromatin accessibility primes gene activity**
171 To better understand the biological basis of GrID-Net's superior accuracy, we applied it on a
172 single-cell multimodal dataset profiling human corticogenesis. We then honed in on peak–gene
173 pairs where GrID-Net and correlation-based methods disagree. Specifically, peak–gene pairs
174 were ranked according to the difference between their GrID-Net and correlation-based
175 prioritization scores (**Methods**), with higher ranking pairs being the ones that were more
176 strongly prioritized by GrID-Net. We then investigated the distinguishing biological
177 characteristics of these differentially-prioritized pairs.
178 One of our motivations behind GrID-Net is the observation that regions of chromatin
179 around lineage-determining genes become accessible prior to gene expression, foreshadowing
180 lineage choice[26,29]. GrID-Net was designed to resolve this cell-state parallax between chromatin
181 accessibility and gene expression and we assessed if it was able to do so. We first associated

182 each peak and each gene with the pseudotime corresponding to its highest value to determine
183 the lag associated with each peak-gene pair. For both correlation-based methods, we found that
184 peak–gene pairs differentially prioritized by GrID-Net are marked by significantly longer
185 pseudotime lags between chromatin accessibility and gene expression ($P < 10^{-26}$, Welch's one-
186 tailed *t*-test, **Figure 2c**, **Methods**). This result supports GrID-Net's ability to account for the
187 asynchrony between noncoding regulatory activity and downstream gene expression control
188 that are expected to characterize causal interactions between the two. To explore the regulatory
189 implications of this asynchrony, we obtained data on histone modification states of enhancers in
190 neural progenitor cells and cross-referenced it with ATAC-seq peaks[30]. In peak–gene pairs
191 differentially prioritized by GrID-Net, more poised enhancers (H3K4me1 only) overlapped with
192 the peaks than active enhancers (both H3K4me1 and H3K27ac)[31] (**Figure 2d**, **Methods**). This
193 was not the case for peak–gene pairs differentially prioritized by correlation-based methods.
194 Poised enhancers are regulatory sites that can trigger future gene expression changes but have
195 not yet affected expression[31,32]; in contrast, active enhancers correspond to ongoing
196 transcription[33]. This suggests that resolving cell-state parallax enables the detection of lagged
197 relationships between the activities of enhancers and their target genes. GrID-Net can thus
198 leverage the concept of chromatin regulatory potential[26] to facilitate a deeper understanding of
199 tissue-specific causal regulatory mechanisms.

200 To evaluate the biological significance of capturing these putative causal interactions, we
201 examined the functions of genes implicated in peak–gene pairs differentially prioritized by GrID-
202 Net. We found these genes to be highly enriched for gene sets associated with neurogenesis
203 regulation and neural progenitor cells ($FDR < 10^{-7}$, GSEA Preranked[34]), directly relating to the
204 process of corticogenesis profiled in this dataset. We emphasize that the enrichment score is
205 calculated only on the peak–gene pairs that GrID-Net prioritizes over the correlation-based
206 methods. In comparison, the set of genes from peak–gene pairs that *all* methods score highly is
207 functionally enriched for housekeeping genes rather than neuronal activity, suggesting that

208 GrID-Net's ability to detect poised enhancers is key to discovering cell-type specific regulatory
209 relationships. Gene expression control during differentiation has been reported to depend highly
210 on asynchronous enhancer-gene interactions[35]. The enrichment of these differentiation-related
211 gene sets thus underscores GrID-Net's ability to construct accurate genome-wide maps of
212 noncoding locus–gene links that go beyond correlative analyses and towards causality.
213

214 **Linking variants to schizophrenia disease genes**

215 The ability to link noncoding loci to the genes they regulate enables the discovery of molecular
216 mechanisms underlying complex diseases. Genome-wide association studies (GWAS) have
217 identified thousands of loci associated with diseases[36]. However, over 90% of GWAS variants
218 lie in noncoding regions[37], and interpreting the functions of these noncoding variants involves
219 the challenge of linking them to the specific genes that they regulate in a cell type-specific
220 manner. With GrID-Net, we can leverage data from a single-cell assay on matching tissue to
221 uncover these cell-type-specific variant-gene links and help bridge the gap between genetic
222 variants and gene dysregulation in disease.

223 We used GrID-Net to study variants associated with schizophrenia (SCZ), a psychiatric
224 disorder for which inherited genetic variants are believed to underlie its pathogenesis[38]. SCZ's
225 high prevalence of around 1%, its substantial morbidity and mortality, and the lack of effective
226 treatments are largely due to our limited understanding of SCZ pathophysiology[39,40]. We applied
227 the GrID-Net peak-gene links inferred from human corticogenesis[27], to interpret and analyze
228 variants reported in the largest GWAS of SCZ to date[41]. We refer to this original study as the
229 Trubetskoy, Pardiñas *et al*. study.

230 To boost GrID-Net's accuracy on shorter-range peak–gene interactions, we incorporated
231 one-dimensional genomic distance as an additional feature for predicting peak–gene
232 associations. We observed a higher likelihood of peak–gene interaction if the two are located
233 close to each other, in line with reported patterns of genomic distance-dependent regulatory

234 effects[42,43]. We calibrated the relative importance of this feature on a dataset profiling mouse
235 skin differentiation[26] where we fitted a logistic regression model that combines the raw GrID-Net
236 score with a genomic distance score (**Methods**). We found that integrating genomic distance
237 with GrID-Net improved predictions of peak–gene links for mouse skin cells. The discriminative
238 power on short-range peak–gene interactions improved, while the ability to detect long-range
239 interactions was preserved. Furthermore, this model trained on mouse skin cell data performed
240 well across species and cell types in human corticogenesis as well (**Figure S3**). This genomic
241 distance-dependent model's scores were used for all remaining analyses of the human
242 corticogenesis dataset.

243 From the SCZ GWAS study, we retrieved all fine-mapped genetic variants with posterior
244 probability (*PP*) over 5%. We intersected these variants with peaks from the human
245 corticogenesis dataset. We then generated an initial set of candidate SCZ genes by associating
246 each variant with up to 3 of its overlapping peak's highest scoring genes, retaining only variant–
247 gene associations with scores in the top 50% of all peak–gene scores. This approach yielded a
248 set of 132 genes that not only is enriched for genes implicated in the original SCZ GWAS study
249 (*FDR* = 0.037, Enrichr[44]) but also includes novel SCZ gene candidates (**Supplemental Table**
250 **1**). For example, of the 12 SCZ variants that were prioritized by both GrID-Net and Trubetskoy,
251 Pardiñas *et al*., 9 were linked to the same target genes by both approaches (**Figure 3a)**. This
252 overlap included the association between rs324017 and NAB2, a gene implicated in stress-
253 induced neuroprotective maintenance that also is the downstream target of the SCZ drug
254 cinnarizine[45,46]. For these same 12 SCZ variants, GrID-Net also proposed 20 additional SCZ
255 candidate genes, of which 12 (60%) were found to be specifically expressed in brain tissues and
256 thus potentially relevant in SCZ etiology (**Methods**). Some of these additional candidates
257 include NXPH4, a neurexophilin involved in synapse function and cerebellar motor control[47],
258 and EPN2, an epsin linked to neural fate commitment[48]. Moreover, of the 78 SCZ variants with
259 *PP* over 50%, only 22 (28%) were linked to genes by the original SCZ GWAS study (**Figure 3b**).

260 Inclusion of SCZ variants that were linked to genes by GrID-Net increased this total to 30 (38%)
261 SCZ variants. GrID-Net can, therefore, be used to complement standard strategies for linking
262 variants to genes, such as those used in the original SCZ GWAS study. It can propose gene
263 associations for variants with no annotations and also augment existing annotations by
264 suggesting additional genes that a variant may be affecting.

265 In particular, GrID-Net helps address a key challenge in GWAS analyses: the common
266 strategy of assigning GWAS variants to their closest gene[49] misses many biologically
267 meaningful variant–gene relationships. By leveraging single-cell heterogeneity, GrID-Net can
268 evaluate longer range interactions. Many of the variant–gene links we identify are characterized
269 by large genomic distances between the variant and the transcription start site (TSS) of the
270 linked gene. Moreover, the variants are often linked to genes that are multiple genes over from
271 their closest gene (**Figure 3c**). For example, DCP1B and NCKAP5L, each of which are
272 connected to only a single variant, are separated from their respective variants by 18 and 11
273 genes (235 and 242 kilobases (kb)), respectively. DCP1B has been implicated in SCZ,
274 hypomania, and memory performance[50,51], while NCKAP5L, a cytoskeletal gene involved in
275 neuronal activity[52], has previously been associated with SCZ as well as amyotrophic lateral
276 sclerosis[53], indicating possible roles for these genes in neuron-specific dysregulation in SCZ.

277 In addition, 52 candidate genes (39.4% of the overall set) were associated with more
278 than one variant (**Figure 3d**). These genes may be controlled by the coordinated activity of
279 multiple regulatory elements and co-regulators relevant to SCZ[54]. IGSF9B was linked to 7
280 noncoding SCZ variants – the most of any gene. Coding variants in IGSF9B have been found to
281 be associated with major depression and the negative symptoms of SCZ[55,56], underscoring its
282 role in regulating affective symptoms in SCZ. Similarly, KCNG2, a potassium channel-encoding
283 gene, was linked to 6 noncoding SCZ variants, and has been proposed as a drug target for SCZ
284 given the known disruption of KCN-related pathways in SCZ[40,57,58]. Neither of these genes were
285 prioritized in the original GWAS analysis.

286 We also assessed the cell-type specificity of the genes linked to SCZ variants. We found
287 that 104 candidate SCZ genes (78.8% of the overall set) are differentially expressed across the
288 cell types profiled in the human corticogenesis dataset (Welch's *t*-test with Benjamini-Hochberg
289 correction $P < 0.05$, **Figure 3e**). Further analysis of the SCZ variants also revealed that 59% of
290 the variants lie in intronic regions, while only 7% of variants lie in intergenic regions more than 5
291 kb from any gene (**Figure 3f**, **Methods**). As enhancers with tissue-specific effects are highly
292 enriched in introns while intergenic enhancers more often regulate housekeeping genes[59], these
293 findings support the potential cell type-specificity of the regulatory disruptions underlying SCZ
294 pathogenesis.
295
296 **Transcription factor binding disruptions are implicated in schizophrenia**
297 Transcription factor (TF) binding at the cis-regulatory elements located in noncoding genomic
298 loci is a key mechanism of gene regulation. We observed that SCZ variants with high *PP* are
299 associated with larger TF binding affinity changes (**Figure S4a**). To investigate the
300 dysregulatory role of SCZ variants prioritized by GrID-Net, we therefore evaluated their effect on
301 TF binding. For each variant associated with a gene by GrID-Net, we calculated the difference
302 in the predicted TF binding affinity between the reference allele and the SCZ risk allele
303 (**Methods**, **Supplemental Table 2**). Of the 115 such SCZ variants, 28 (these were associated
304 with 49 target genes) were predicted to cause a 20-fold or stronger change in at least one TF's
305 binding affinity (**Figure 4a**). The TFs we identified as having the most widespread binding
306 affinity changes are known to be implicated in SCZ, exemplifying GrID-Net's ability to deliver de
307 novo mechanistic insight. VEZF1, a vasculature-related TF reported to play a role in SCZ
308 dysregulation[60], displayed diminished binding affinity most frequently (in 5 variants). In addition,
309 KLF15 and KLF6, two Kruppel-like family (KLF) TFs known to regulate neuronal growth and
310 linked to SCZ[61,41], were each associated with reduced binding affinity in 4 variants (**Figure S4b**).
311 Meanwhile, the TFs with the most widespread increase in binding affinity – SP1, SP3, and

312 KLF16 – have also been reported to be disrupted in SCZ[62,63] (**Figure S4c**). More broadly, of the
313 TFs with a 20-fold binding affinity change in at least one variant, 57.9% (=33/57) have cell type-
314 specific expression patterns in corticogenesis (Welch's *t*-test with Benjamini-Hochberg
315 correction $P < 0.05$). Collectively, these results point to genetic disruption of neuron-specific
316 regulatory elements as potential drivers of SCZ pathogenesis.

317 We reasoned that if the TF binding affinity changes at a SCZ variant are phenotypically
318 meaningful, they should result in expression changes of the corresponding gene(s) (as identified
319 by GrID-Net). Accordingly, we obtained gene expression data from glial progenitor cells (GPCs)
320 produced from SCZ patient-derived induced pluripotent cells, as well as matching controls[58,64],
321 and computed differential expression patterns in SCZ GPCs. Focusing on glial differentiation
322 genes, we observed that the differential expression of a gene in SCZ GPCs was indeed
323 significantly correlated with TF binding affinity changes at the corresponding SCZ variant
324 (Pearson correlation coefficient = 0.59, **Figure 4b**, **Methods**). The consistency between these
325 independent observations provides support to our overall analytic approach and underscores
326 the potential to leverage GrID-Net for examining causal, gene-specific regulatory mechanisms.

327 We drilled down into specific genes reported by our analysis, focusing on those linked to
328 SCZ variants associated with at least 20-fold TF binding affinity changes (**Figure 4c**). NAB2, the
329 aforementioned gene implicated in neuroprotective processes[46], was linked to rs324017, which
330 contributes to a 200-fold decline in binding affinity for the TF PURA. Intriguingly, PURA is also
331 critical for neural function and development, and its dysfunction is implicated in several
332 neurological and cognitive impairments[65,66]. In addition, rs12991836 was linked to ZEB2, whose
333 TSS is positioned 137 kb from the variant. ZEB2 has been previously reported as a target of
334 long-range regulation, and heterozygous ZEB2 mutations are known to cause Mowat-Wilson
335 syndrome, which features structural brain abnormalities and intellectual disability[67].

336 Furthermore, GrID-Net prioritized two potassium transport genes - KCNG2 and
337 SLC12A6 - a class of genes which has been associated with compromised potassium uptake

338 when dysregulated[58]. This phenotype is characteristic of not only SCZ but also other neuron-
339 related diseases, such as Huntington's disease, multiple sclerosis, and hemiplegic migraine[64,68].
340 KCNG2 was linked to 6 SCZ variants, the second most of any gene as described earlier. Using
341 TF binding affinity changes to hone in on the most relevant variants, we identified rs8091497 as
342 the variant with the largest TF binding affinity change. rs8091497 was associated with a 40-fold
343 decline in binding affinity for SP1, which itself has been linked to behavioral abnormalities in
344 SCZ[63]. As for SLC12A6, its corresponding variant (rs117799466) increases the binding affinity
345 for many TFs with repressive capabilities. The three TFs with the greatest binding affinity gains
346 at this locus – SP1, ZN281, and MAZ – all play a role in neuronal differentiation[63,69,70], and both
347 SP1 and ZN281 have also shown increased expression levels in SCZ as well[71,72]. Combined
348 with their increased binding affinity incurred by rs117799466, the upregulation of these TFs may
349 contribute to the repression of SLC12A6, an outcome that has been previously linked to the
350 potassium uptake impairments that characterize SCZ glia[58]. Together, these findings suggest a
351 possible mechanism for SLC12A6 downregulation in SCZ that implicates enhanced binding of
352 neuron-related repressors to a SLC12A6-targeting enhancer to reduce its expression levels.
353 We then broadened our analysis to gene modules, exploring if the genes identified by
354 GrID-Net are systematically involved in SCZ-related pathways. The 14 SCZ variants that cause
355 at least 40-fold changes in TF binding affinity are linked by GrID-Net to 26 genes, whose
356 differential expression patterns we examined in SCZ GPCs. We detected 3 distinct gene
357 modules that are co-expressed with several of these genes (**Figure 4d**, **Methods**). Module A
358 (SLC12A6, DLX1, and DLX2) is enriched for genes related to noradrenergic neuron
359 differentiation ($FDR < 0.05$) and morphogenesis ($FDR < 0.02$); Module B (MFAP4, STAT6,
360 NXPH4) is enriched for extracellular matrix organization ($FDR < 10^{-14}$) and angiogenesis
361 regulation ($FDR < 0.02$) genes, and Module C (EPN2, P2RX7, ICA1) is enriched for synaptic
362 transmission ($FDR < 10^{-4}$) and postsynaptic membrane potential ($FDR < 0.05$) (Enrichr[44]). Each
363 of these pathways has been previously associated with SCZ. Noradrenergic neurons have long

364 been implicated in disrupted neuromodulation in SCZ[73], neuronal angiogenesis underlies the
365 vascular abnormalities observed in SCZ[74], and genetic studies have pointed to the critical role of
366 synaptic function in SCZ[41]. These pathways, therefore, enable the interpretation of SCZ variants
367 beyond the dysregulation of the specific genes that they directly affect, presenting an
368 opportunity to discover systems-level mechanisms of SCZ pathogenesis.
369

370 **Discussion**

371 Most methods for single-cell multimodal data integration have focused on synchronizing
372 modalities to improve downstream tasks, such as cell type identification. In contrast, we
373 demonstrate it is also powerful to systematically study the asynchrony between the modalities.
374 We presented GrID-Net, a method to uncover causal connections between genes and
375 noncoding genomic loci using co-assayed single-cell RNA-seq and ATAC-seq data. GrID-Net
376 operates at a genome-wide scale, is tissue-specific, and can effectively detect long-range (up to
377 1 Mb) cis-regulatory associations. A longstanding challenge with GWAS studies has been to
378 identify the specific genes that are dysregulated by noncoding disease-linked variants. With
379 GrID-Net, a single commercially-available assay of the matching tissue can power a genome-
380 wide estimation of variant-gene relationships. By resolving tissue-specific causal variant-gene
381 associations, GrID-Net can help identify disease-linked genes as therapeutic targets.
382

383 When studying a particular disease, GrID-Net should be paired with a single-cell assay of a
384 matching tissue undergoing a dynamic biological process, such as development, differentiation,
385 disease, or drug response. GrID-Net can then translate the results of this single assay to
386 genome-wide causal hypotheses specific to the disease. We demonstrate this by leveraging a
387 single-cell assay of human corticogenesis to interpret the functional effects of schizophrenia
388 (SCZ) variants. GrID-Net expands both the breadth and depth of existing gene prioritization
389 analyses: it finds novel associations for many unlinked variants while also uncovering additional

390 causal mechanisms for previously investigated variants. For instance, we identified SCZ
391 variants causing large (>20-fold) changes in TF binding affinity and connected them to 49
392 genes. Many of these genes are located distally from the SCZ variants that are predicted to
393 affect them and would have been neglected by existing GWAS variant–gene associations
394 analyses that consider only genomic proximity. Furthermore, by correlating TF binding affinity
395 changes with differential expression in SCZ glial progenitor cells, we predicted possible
396 mechanisms of TF-driven dysregulation for genes like KCNG2 and SLC12A6. Combined with an
397 exploration of SCZ-related pathways associated with GrID-Net-identified genes, these analyses
398 provide a comprehensive strategy for unveiling mechanistic insights about SCZ pathogenesis.
399 Such analyses can also be applied to investigate other diseases.

400 Our work rests on two conceptual insights that are broadly applicable. First, when
401 different modalities are measured in the same cell their causal relationship implies a temporal
402 offset between them, a phenomenon we refer to as “cell-state parallax”. Here, we exploit the
403 parallax between RNA-seq and ATAC-seq, but such parallax can also be leveraged from other
404 co-assayed modalities such as chromatin structure or methylation. We view RNA velocity[75] and
405 protein velocity[76] as specific applications of the broader concept of cell-state parallax, the
406 distinction being that velocity uses pre-defined mappings between modalities (e.g., spliced vs.
407 unspliced transcript-counts for a particular gene) to infer the differentiation direction of each cell.
408 In contrast, such mappings between modalities may be unclear in general cell-state parallax
409 and require customized inference approaches. Accordingly, our second conceptual insight is to
410 develop computational frameworks that can accommodate the cross-modality dynamics of cell
411 differentiation. Much of the existing work on modeling gene expression dynamics utilizes
412 mathematical and statistical tools designed for time-series analysis[77,78], which make the
413 unrealistic assumption that a total ordering of cells can capture the underlying biology. We
414 argue that a partial ordering of cells is necessary to accurately model forks and merges in
415 trajectories and accordingly formulate the differentiation landscape as a directed acyclic graph

of cells that can represent such an ordering. We expect a promising direction of future research will be computational inference techniques, like GrID-Net, that can operate under this more realistic model.

Since GrID-Net operates on observational data, the locus-gene associations it infers are causal hypotheses that need to be validated by perturbations. In the eQTL-based evaluation, we sought to mimic such validations by investigating random genetic variants observed in the population and cross-referencing them to observed expression changes. We also note that such causal interactions are likely indirect and mediated by additional biophysical events, such as transcription factor binding. Nonetheless, indirect causality is biologically and therapeutically valuable, as the underlying mechanisms can still be exploited via interventions to achieve the desired phenotype.

With the growing availability of single-cell multimodal datasets, our approach is positioned to serve as a key framework for unlocking new mechanisms of gene regulatory dynamics that underlie fundamental biology and disease. Moreover, GrID-Net can be applied broadly to other domains featuring graph-based dynamics, such as uncovering the drivers of social media disinformation cascades. A key technical advance of this work is generalizing the applicability of standard Granger causal analysis to a wider set of domains than previously possible. Doing so required the use of graph-theoretic deep learning techniques to implement the core statistical intuition of Granger causality for any DAG-structured ordering of observations. Our work, like others before us[79], thus highlights how classical statistical concepts can be reimagined with modern deep learning techniques to dramatically widen their applicability.

440 **Methods**

441

442 **GrID-Net formulation**

443 See Supplementary Note.

444

445 **GrID-Net training details and hyperparameters**

446 All GrID-Net models consisted of $L = 10$ GNN layers and were trained using the Adam optimizer
447 with a learning rate of 0.001 and a mini-batch size of 1024 candidate peak-gene pairs. Models
448 were trained for a maximum of 20 epochs or until the relative change in the loss function was
449 less than $0.1/|P|$ between successive epochs. Here, $|P|$ denotes the number of candidate peak-
450 gene pairs evaluated per dataset. GrID-Net models were implemented in PyTorch and trained
451 on a single NVIDIA Tesla V100 GPU.

452

453 **Single-cell multimodal datasets**

454 We used four single-cell multimodal datasets that spanned a range of different multimodal
455 profiling techniques: 1) a 10x Multiome dataset profiling human corticogenesis, 2) a SHARE-seq
456 dataset profiling mouse skin differentiation, 3) a SNARE-seq dataset profiling mouse neonatal
457 cerebral cortical cells, 4) a sci-CAR dataset profiling A549 cells treated with dexamethasone.
458 For each of these datasets, we followed recommended practice[80] and applied $log\ (1 +$
459 $CPM/100)$ and $log\ (1 + CPM/10)$ transformations to the RNA-seq and ATAC-seq counts data,
460 respectively, where CPM indicates counts per million. We then filtered out all genes that were
461 represented in fewer than 1% of cells in the dataset and all peaks that were represented in
462 fewer than 0.1% of cells. We assembled a set of candidate peak-gene pairs by selecting all
463 peaks with centers within 1 Mb from each gene's TSS. For the 10x Multiome dataset, we only
464 considered peak-gene pairs that were separated by 250 kb to cohere with the genomic distance

465 threshold used in the original study[27]. To consider multiple samples for each dataset, we
466 generated 5 uniformly sampled sets of 5,000 cells for each of the 10x Multiome, SHARE-seq,
467 and SNARE-seq datasets, while we generated 5 uniformly sampled sets of 3,000 cells for the
468 sci-CAR dataset due its smaller quantity of cells (3,081 cells).
469
470 **Directed acyclic graph (DAG) construction**
471 For each sample of a dataset, we first generated per-cell representations for each of the gene
472 expression and chromatin accessibility modalities. Gene expression representations were
473 constructed by (1) scaling the transformed RNA-seq data to unit variance and zero mean, (2)
474 retaining the top 2000 most highly variable genes, (3) and projecting these normalized
475 expression values onto their top 100 principal components (excluding components that have
476 Spearman correlation $\rho > 0.9$ with the total counts per cell. Chromatin accessibility
477 representations were obtained transforming the ATAC-seq data using tf-idf[81,82] and projecting
478 the resulting values onto the top principal components as in (3) above. We then used Schema[20]
479 to learn cell representations that unify information from these gene expression and chromatin
480 accessibility representations. We constructed a k-nearest neighbor graph ($k = 15$) based on the
481 Euclidean distances between these cell representations and inferred pseudotime using the
482 scanpy[83] implementation of diffusion pseudotime[84] with default parameters. We then removed
483 edges in the k-nearest neighbor graph if they did not align with the direction of increasing
484 pseudotime, thus ensuring that no cycles exist and the resulting graph is a DAG.
485
486 **Alternative methods**
487 *Pearson correlation*. We computed the Pearson correlation between the transformed ATAC-seq
488 and RNA-seq values for each candidate peak-gene pair, and the absolute value of the
489 correlation coefficients were used to score the peak-gene pairs.
490

491 *Pseudocell correlation*. For each sample, we identified the 49 nearest neighbors for each cell
492 based on the Euclidean distances between the joint representations learned from Schema (see
493 "Directed acyclic graph construction"). We constructed a "pseudocell" from each cell and its 49
494 nearest neighbors, constructing its RNA-seq and ATAC-seq profiles by summing the raw counts
495 from each of the modalities for all cells included in the pseudocell. We then applied $log\ (1 +$
496 $CPM/100)$ and $log\ (1 + CPM/10)$ transformations to the RNA-seq and ATAC-seq counts data,
497 respectively. Pearson correlation was applied to the transformed ATAC-seq and RNA-seq
498 values for each candidate peak-gene pair, and the absolute value of the correlation coefficients
499 were used to score the peak-gene pairs.
500
501 *Activity-by-contact (ABC) model*. We obtained ABC prediction scores in cell types matching the
502 cell types profiled in the single-cell multimodal studies. For each single-cell study, we
503 intersected the candidate peak-gene pairs with the ABC enhancer-gene pairs, assigning each
504 peak-gene pair the corresponding ABC score. If no ABC enhancer-gene pair overlap was
505 identified, the peak-gene pair was assigned a value of 0, as the ABC prediction scores were
506 available only for enhancer-gene pairs that exceeded the threshold of 0.015 defined in the
507 original study[10]. For single-cell studies that profiled mouse cells, we mapped ABC enhancer-
508 gene pairs in matching human cell types to the mouse genome due to the strong conservation
509 of epigenomes between human and mouse[85,86].
510
511 **eQTL evaluations**
512 We obtained eQTL data from the eQTL Catalogue[87] and used eQTL data specific to the cell
513 type profiled in each single-cell multimodal study. For mouse tissues, we mapped eQTL results
514 for matching human tissues to the mouse genome, as human and mouse epigenomes show
515 strong conservation[85,86].

516 We first intersected eQTL variant-gene pairs with peak-gene pairs from the single-cell
517 data and assigned each remaining peak-gene pair the most statistically significant eQTL. We
518 then binned these peak-gene pairs according to the genomic distance between each gene's
519 TSS and each peak's center, creating a bin for each of the following ranges of genomic
520 distances: [1 bp, 1 kb), [1 kb, 10kb), [10kb, 100kb), [100kb, 1 Mb). We identified all peak-gene
521 pairs that overlapped significant eQTL variant-gene pairs ($P < 0.001$), and these peak-gene
522 pairs served as the positive examples. To generate a distance-matched set of negative
523 examples, we determined the number of positive examples in each genomic distance bin and
524 selected an equal number of non-significant eQTL variant-gene pairs with the highest *P* values
525 for that bin.
526
527 **Chromatin interactions evaluations**
528 We obtained Hi-C data for cell types matching the 10x Multiome, SNARE-seq, and sci-CAR
529 datasets[88–90]. We used 5C data for the SHARE-seq dataset[91] due to the unavailability of mouse
530 skin Hi-C data. Chromatin interactions were mapped to a peak-gene pair by intersecting the
531 genomic windows in the chromatin interactions data with the peak and the TSS of the gene. For
532 Hi-C data, peak-gene pairs that were associated with the top 1% of chromatin contact
533 frequencies were treated as positive examples, while the remaining peak-gene pairs were
534 considered negative examples. For the 5C dataset, which evaluated the statistical significance
535 of chromatin loops, peak-gene pairs that were associated with significant looping interactions (P
536 < 0.001) were treated as positive examples, and the remaining peak-gene pairs were
537 considered negative examples.
538
539 **GrID-Net vs. correlation peak-gene prioritization**
540 For the human corticogenesis dataset, we averaged the rankings of all candidate peak-gene
541 pairs across the five trials for each of the GrID-Net, Pearson correlation, and pseudocell

correlation methods. We then calculated the difference between GrID-Net's average rankings for each peak-gene pair with those of the Pearson correlation and pseudocell correlation methods. The peak-gene pairs with the most positive differences represent the peak-gene pairs that GrID-Net prioritizes over the method against which GrID-Net is being compared. To evaluate enrichment of gene sets in peak-gene pairs prioritized by GrID-Net, genes were ranked with respect to these differences and enrichment was computed using GSEA Preranked[34].

Enhancer states

We obtained H3K27ac and H3K4me1 ChIP-seq data for all samples of human neural progenitor cells and fetal neural cells in ChIP-Atlas [30]. We identified statistically significant ChIP-seq peaks (MACS2 $Q < 10^{-10}$) that were positioned fully within ATAC-seq peaks from the human corticogenesis dataset. Active enhancers were defined as ATAC-seq peaks containing overlapping H3K4me1 and H3K27ac ChIP-seq peaks, while poised enhancers were defined as ATAC-seq peaks containing H3k4me1 ChIP-seq peaks and no H3K27ac ChIP-seq peaks.

Genomic distance-dependent GrID-Net scores

We constructed a general model that integrated GrID-Net scores with genomic distance to account for known patterns of genomic distance-dependent transcription factor (TF) regulatory effects[42,43]. The model assumes that TF regulatory effects exponentially decay with increasing genomic distance. To learn the decay rate, we trained a logistic regression model to predict peak-gene pairs overlapping significant skin eQTLs ($P < 10^{-3}$) for the SHARE-seq mouse skin dataset.

$$\mathrm{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \beta_1 \log(s_i) + \beta_2 \hat{d}_i$$

Here, π_i is the probability of peak-gene pair i being a positive example, s_i corresponds to the mean rankings of the GrID-Net scores across trials for the peak-gene pair, and $\widehat{d_i} =$

567 $max(log_{10}(d_i) - 4,0)$, where d_i is the genomic distance in base pairs (bp) between the center
568 of the peak and the TSS of the gene in the peak-gene pair. This formulation assumes that only
569 distal interactions at genomic distances greater than 10 kb are affected by the genome
570 distance-dependent TF regulatory effect.
571 Positive and negative examples were chosen as described in "eQTL evaluations". We
572 held out one-third of the example as a test set, and then performed nested 3-fold cross
573 validation on the remaining examples, selecting the model that yielded the highest AUROC
574 across the folds. This model increased the AUROC on the test set compared to the predictions
575 from the averaged GrID-Net rankings (**Figure S3a**). We applied this pre-trained model on the
576 human corticogenesis dataset and found that this model outperformed the baseline predictions
577 from the averaged GrID-Net rankings (**Figure S3b**). These results demonstrated the
578 effectiveness of incorporating genomic distance as an additional feature alongside scores
579 learned from the single-cell multimodal data.
580
581 **Linking SCZ GWAS Variants to Genes**
582 We obtained FINEMAP posterior probabilities (*PP*) calculated in the original SCZ GWAS
583 study[41]. We retained all fine-mapped variants with $PP \geq 0.05$ that overlapped peaks from the
584 human corticogenesis dataset. We identified the peak that matched each SCZ variant and,
585 using the genomic distance-dependent GrID-Net scores, we linked the variant to up to 3 of the
586 peak's highest scoring genes, retaining only peak-gene pairs in the top 50% of all peak-gene
587 scores.
588
589 **Brain-specific gene expression**
590 We obtained consensus tissue-specific transcript levels from the Human Protein Atlas[92,93]. For a
591 given gene, we first calculated the mean and standard deviation of the gene's transcript levels

592 across all non-brain tissues. The brains-specific tissues were identified to be those from the
593 basal ganglia, cerebellum, cerebral cortex, medulla oblongata, midbrain, pons, and white
594 matter. We then converted the transcript level for the gene in each of the brain-specific tissues
595 into a z-score with respect to the non-brain tissues ($z = \frac{x-\mu_{non-brain}}{\sigma_{non-brain}}$, where x is the transcript
596 level for the gene in a brain-specific tissue). Genes that had a z-score over 1 in any of the brain-
597 specific tissues were annotated as having brain-specific gene expression.
598
599 **TF binding affinity changes**
600 We obtained the reference and SCZ risk alleles for all SCZ variants from the original SCZ
601 GWAS study[41]. We then flanked each of the alleles with the genomic sequences $\pm$20 bp of the
602 SCZ variant position using the GRCh38/hg38 reference genome[94], generating a reference
603 sequence and SCZ variant sequence. We applied FIMO[95] to each of these sequences to predict
604 TF binding affinities using position weight matrices (PWMs) from HOCOMOCO v11 (Full
605 Collection)[96], JASPAR CORE (Non-Redundant) v2[97], and CIS-BP v2[98]. We evaluated the effect
606 of a specific SCZ variant on a TF's binding affinity by comparing the FIMO-calculated *p*-values
607 for the TF between the reference sequence and SCZ variant sequence.
608
609 **SCZ GPC gene expression**
610 We obtained bulk RNA-seq data from glial progenitor cells (GPCs) produced from 4 SCZ
611 patient-derived induced pluripotent cells (iPSCs) and paired controls subejcts[58,64], with at least 3
612 samples per patient. We applied a $log\ (1 + CPM)$ transformation to this data.
613 To evaluate the relationship between TF binding affinity changes caused by SCZ
614 variants to expression changes in GPCs, we first determined peak-gene pairs that are specific
615 to glial differentiation. In the human corticogenesis dataset, we identified genes that were
616 differentially expressed between glial cells and the cycling progenitor cells, which were the root

617 of the corticogenesis trajectory ($P < 0.05$, Welch's *t*-test). We then used Welch's *t*-test to
618 compare the expression levels of these glial differentiation-specific genes for each SCZ patient
619 against all of the control samples. For each gene, we identified the most statistically significant
620 difference so as to account for all genes that may be differentially expressed in any of the
621 patients compared to the controls.
622
623 **SCZ GPC co-expression pathways**
624 We first identified a set of 26 genes linked to SCZ variants causing at least a 40-fold change in
625 binding affinity for any TF. We identified 3 sets of these SCZ genes that were correlated with
626 each other (Pearson correlation coefficient ≥ 0.6) in the SCZ GPCs and paired controls
627 (SLC12A6, DLX1, DLX2; MFAP4, STAT6, NXPH4; EPN2, P2RX7, ICA1). For each of these
628 sets of SCZ genes, we then determined gene modules that may be implicated in their
629 dysregulation by identifying genes that are highly co-expressed (Pearson correlation coefficient
630 ≥ 0.8) with any of the SCZ genes in the set. We applied gene set enrichment analysis using
631 Enrichr to detect biological processes that were enriched in each of these modules.
632

References

1. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535–1548.e16 (2018).
2. Papp, B. & Plath, K. Epigenetics of Reprogramming to Induced Pluripotency. *Cell* vol. 152 1324–1343 Preprint at https://doi.org/10.1016/j.cell.2013.02.043 (2013).
3. Pierce, S. E., Granja, J. M. & Greenleaf, W. J. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.* **12**, 2969 (2021).
4. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
5. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* vol. 16 197–212 Preprint at https://doi.org/10.1038/nrg3891 (2015).
6. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 368 20120362 Preprint at https://doi.org/10.1098/rstb.2012.0362 (2013).
7. Shukla, A. & Huangfu, D. Decoding the noncoding genome via large-scale CRISPR screens. *Curr. Opin. Genet. Dev.* **52**, 70–76 (2018).
8. Wang, Q. *et al.* A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* **22**, 691–699 (2019).
9. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
10. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* vol. 593 238–243 Preprint at https://doi.org/10.1038/s41586-021-03446-x (2021).
11. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of

single cells. *Science* **361**, 1380–1385 (2018).

12. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
13. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
14. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nat. Methods* **17**, 14–17 (2020).
15. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).
16. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
17. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
18. Kamimoto, K., Hoffmann, C. M. & Morris, S. A. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv* 2020.02.17.947416 (2020) doi:10.1101/2020.02.17.947416.
19. Li, C., Virgilio, M. C., Collins, K. L. & Welch, J. D. Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nat. Biotechnol.* **41**, 387–398 (2022).
20. Singh, R., Hie, B. L., Narayan, A. & Berger, B. Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome Biol.* **22**, 131 (2021).
21. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).
22. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).

23. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).

24. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* vol. 37 424 Preprint at https://doi.org/10.2307/1912791 (1969).

25. Shojaie, A. & Fox, E. B. Granger Causality: A Review and Recent Advances. *Annual Review of Statistics and Its Application* vol. 9 289–319 Preprint at https://doi.org/10.1146/annurev-statistics-040120-010930 (2022).

26. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).

27. Trevino, A. E. *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069.e23 (2021).

28. Singh, R. & Berger, B. Deciphering the species-level structure of topologically associating domains. *bioRxiv* 2021.10.28.466333 (2021) doi:10.1101/2021.10.28.466333.

29. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).

30. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac199.

31. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).

32. Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).

33. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol.* **22**, 108 (2021).

34. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for

interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

35. Herrmann, J. C., Beagrie, R. A. & Hughes, J. R. Making connections: enhancers in cellular differentiation. *Trends Genet.* **38**, 395–408 (2022).
36. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
37. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
38. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
39. Malan-Müller, S. *et al.* A systematic review of genetic variants associated with metabolic syndrome in patients with schizophrenia. *Schizophr. Res.* **170**, 1–17 (2016).
40. Lago, S. G. & Bahn, S. The druggable schizophrenia genome: from repurposing opportunities to unexplored drug targets. *NPJ Genom Med* **7**, 25 (2022).
41. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
42. Chen, C.-H. *et al.* Determinants of transcription factor regulatory range. *Nat. Commun.* **11**, 2472 (2020).
43. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* **8**, 2502–2515 (2013).
44. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).
45. Karunakaran, K. B., Chaparala, S. & Ganapathiraju, M. K. Potentially repurposable drugs for schizophrenia identified from its interactome. *Sci. Rep.* **9**, 12682 (2019).
46. Marballi, K. K. & Gallitano, A. L. Immediate Early Genes Anchor a Biological Pathway of Proteins Required for Memory Formation, Long-Term Depression and Risk for Schizophrenia. *Front. Behav. Neurosci.* **12**, 23 (2018).

47. Meng, X. *et al.* Neurexophilin4 is a selectively expressed α-neurexin ligand that modulates specific cerebellar synapses and motor functions. *Elife* **8**, (2019).

48. Cardano, M. *et al.* Epsins Regulate Mouse Embryonic Stem Cell Exit from Pluripotency and Neural Commitment by Controlling Notch Activation. *Stem Cells Int.* **2019**, 4084351 (2019).

49. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).

50. Barkhuizen, W., Pain, O., Dudbridge, F. & Ronald, A. Genetic overlap between psychotic experiences in the community across age and with psychiatric disorders. *Transl. Psychiatry* **10**, 86 (2020).

51. Schröder, J. *et al.* MicroRNA-138 is a potential regulator of memory performance in humans. *Front. Hum. Neurosci.* **8**, 501 (2014).

52. Chahrour, M. H. *et al.* Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genet.* **8**, e1002635 (2012).

53. McLaughlin, R. L. *et al.* Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nat. Commun.* **8**, 14774 (2017).

54. Kim, J. *et al.* The co-regulation mechanism of transcription factors in the human gene regulatory network. *Nucleic Acids Res.* **40**, 8849–8861 (2012).

55. Shyn, S. I. *et al.* Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol. Psychiatry* **16**, 202–215 (2011).

56. Fabbri, C. & Serretti, A. Role of 108 schizophrenia-associated loci in modulating psychopathological dimensions in schizophrenia and bipolar disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **174**, 757–764 (2017).

57. Devor, A. *et al.* Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia. *Mol. Psychiatry* **22**, 792–801 (2017).

58. Liu, Z. *et al.* Dysregulated Glial Differentiation in Schizophrenia May Be Relieved by

Suppression of SMAD4- and REST-Dependent Signaling. *Cell Rep.* **27**, 3832–3843.e6 (2019).

59. Borsari, B. *et al.* Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res.* **31**, 1325–1336 (2021).
60. Chen, R. *et al.* Cell type–specific mechanism of Setd1a heterozygosity in schizophrenia pathogenesis. *Science Advances* vol. 8 Preprint at https://doi.org/10.1126/sciadv.abm1077 (2022).
61. Moore, D. L., Apara, A. & Goldberg, J. L. Krüppel-like transcription factors in the nervous system: novel players in neurite outgrowth and axon regeneration. *Mol. Cell. Neurosci.* **47**, 233–243 (2011).
62. Peyrot, W. J. & Price, A. L. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. *Nat. Genet.* **53**, 445–454 (2021).
63. Pinacho, R. *et al.* Analysis of Sp transcription factors in the postmortem brain of chronic schizophrenia: a pilot study of relationship to negative symptoms. *J. Psychiatr. Res.* **47**, 926–934 (2013).
64. Seifert, G., Schilling, K. & Steinhäuser, C. Astrocyte dysfunction in neurological disorders: a molecular perspective. *Nat. Rev. Neurosci.* **7**, 194–206 (2006).
65. Tanaka, A. J. *et al.* De novo mutations in PURA are associated with hypotonia and developmental delay. *Cold Spring Harb Mol Case Stud* **1**, a000356 (2015).
66. Daniel, D. C. & Johnson, E. M. PURA, the gene encoding Pur-alpha, member of an ancient nucleic acid-binding protein family with mammalian neurological functions. *Gene* **643**, 133–143 (2018).
67. Barešić, A., Nash, A. J., Dahoun, T., Howes, O. & Lenhard, B. Understanding the genetics of neuropsychiatric disorders: the potential role of genomic regulatory blocks. *Mol. Psychiatry* **25**, 6–18 (2020).
68. Bøttger, P. *et al.* Glutamate-system defects behind psychiatric manifestations in a familial

hemiplegic migraine type 2 disease-mutation mouse model. *Sci. Rep.* **6**, 22047 (2016).

69. Pieraccioli, M. *et al.* ZNF281 inhibits neuronal differentiation and is a prognostic marker for neuroblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7356–7361 (2018).
70. Okamoto, S.-I., Sherman, K., Bai, G. & Lipton, S. A. Effect of the ubiquitous transcription factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal differentiation. *Brain Res. Mol. Brain Res.* **107**, 89–96 (2002).
71. Pinacho, R. *et al.* Increased SP4 and SP1 transcription factor expression in the postmortem hippocampus of chronic schizophrenia. *J. Psychiatr. Res.* **58**, 189–196 (2014).
72. Wen, Y.-D. *et al.* Genetic Profiles Playing Opposite Roles of Pathogenesis in Schizophrenia and Glioma. *J. Oncol.* **2020**, 3656841 (2020).
73. Mäki-Marttunen, V., Andreassen, O. A. & Espeseth, T. The role of norepinephrine in the pathophysiology of schizophrenia. *Neurosci. Biobehav. Rev.* **118**, 298–314 (2020).
74. Maletic, V., Eramo, A., Gwin, K., Offord, S. J. & Duffy, R. A. The Role of Norepinephrine and Its α-Adrenergic Receptors in the Pathophysiology and Treatment of Major Depressive Disorder and Schizophrenia: A Systematic Review. *Front. Psychiatry* **8**, 42 (2017).
75. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
76. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020).
77. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).
78. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Rep.* **38**, 110333 (2022).
79. Andrew, G., Arora, R., Bilmes, J. & Livescu, K. Deep Canonical Correlation Analysis. in *Proceedings of the 30th International Conference on Machine Learning* (eds. Dasgupta, S. & McAllester, D.) vol. 28 1247–1255 (PMLR, 17--19 Jun 2013).
80. Hie, B. *et al.* Computational Methods for Single-Cell RNA Sequencing. *Annual Review of*

Biomedical Data Science vol. 3 339–364 Preprint at https://doi.org/10.1146/annurev-biodatasci-012220-100601 (2020).

81. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
82. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309–1324.e18 (2018).
83. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
84. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
85. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
86. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392 (2012).
87. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
88. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
89. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* vol. 485 376–380 Preprint at https://doi.org/10.1038/nature11082 (2012).
90. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
91. Poterlowicz, K. *et al.* 5C analysis of the Epidermal Differentiation Complex locus reveals distinct chromatin interaction networks between gene-rich and gene-poor TADs in skin epithelial cells. *PLoS Genet.* **13**, e1006966 (2017).

92. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

93. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).

94. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

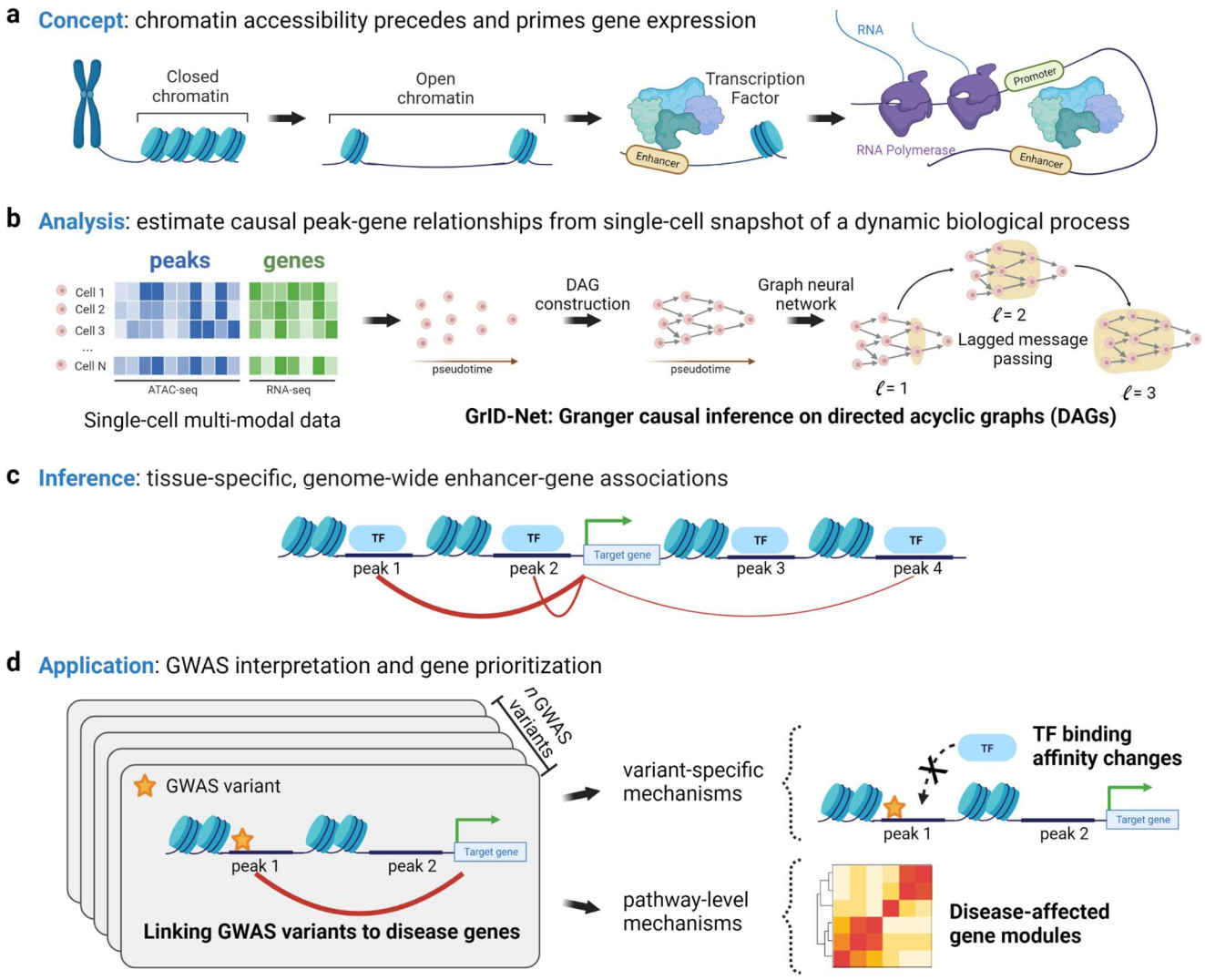
95. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

96. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

97. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* vol. 50 D165–D173 Preprint at https://doi.org/10.1093/nar/gkab1113 (2022).

98. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

633 **Figures**



634
635

636 **Figure 1. GrID-Net Framework. (a)** The causal asynchrony between chromatin accessibility
637 and gene expression results in “cell-state parallax”, where a temporal lag exists between these
638 two modalities. Leveraging this parallax from single-cell snapshots enables inference of causal
639 locus-gene associations. **(b)** To identify links between noncoding loci (i.e. peaks) and the genes
640 they regulate, we apply GrID-Net to single-cell multimodal data in which chromatin accessibility
641 (ATAC-seq) and gene expression (RNA-seq) have been co-assayed in the same cell. GrID-Net
642 is an extension of classical Granger causality that can operate on DAG-based representations

643 of single-cell trajectories. The core of GrID-Net is a graph neural network-based lagged
644 message passing architecture. **(c)** GrID-Net enables the detection of asynchronous interactions
645 between regulatory activity at accessible noncoding loci and their downstream effects on gene
646 expression. **(d)** GrID-Net peak-gene links enable the functional interpretation of genetic variants
647 associated with disease. We present a framework for studying the functional effects of such
648 variants both at the level of specific variants and with respect to disease-affected modules.
649

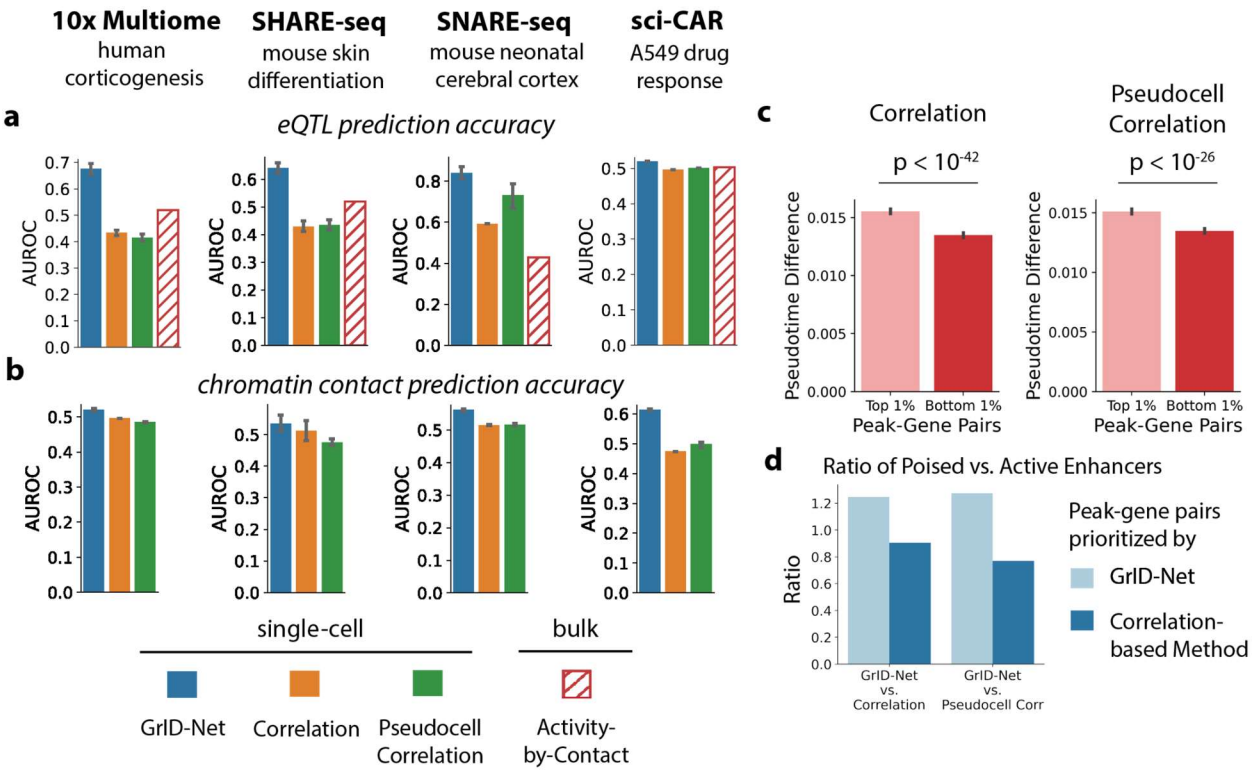


Figure 2. GrID-Net links noncoding loci to genes from single-cell multimodal dynamics. **(a)** Accuracy of single-cell (GrID-Net, Pearson correlation, and pseudocell correlation) and bulk (Activity-By-Contact) methods in predicting expression quantitative trait loci (eQTLs) in matching tissue. In this evaluation, we use eQTLs as proxies for causal interactions between genomic loci and genes. **(b)** Accuracy of methods in predicting high frequency three-dimensional chromatin contacts. The Activity-By-Contact (ABC) model is excluded here since it uses chromatin interaction data as one of its inputs. **(c)** Compared to existing approaches, GrID-Net is better able to identify asynchronous peak-gene associations. On the human corticogenesis dataset, we collected peak-gene pairs that GrID-Net prioritized differently than the correlation-based approaches (top 1%: preferred by GrID-Net; bottom 1%: preferred by the other approach). For each pair, we quantified the asynchrony as the separation between pseudotime values corresponding to the cells where the peak and gene achieved their maximum values,

665 respectively. **(d)** GrID-Net effectively captures instances of chromatin priming. On the human
666 corticogenesis dataset, we show the ratio of poised enhancers to active enhancers for peaks in
667 peak-gene pairs prioritized by GrID-Net vs. each of the correlation-based methods. The
668 estimation of poised-vs.-active enhancers was done using publicly-available histone
669 modification data.

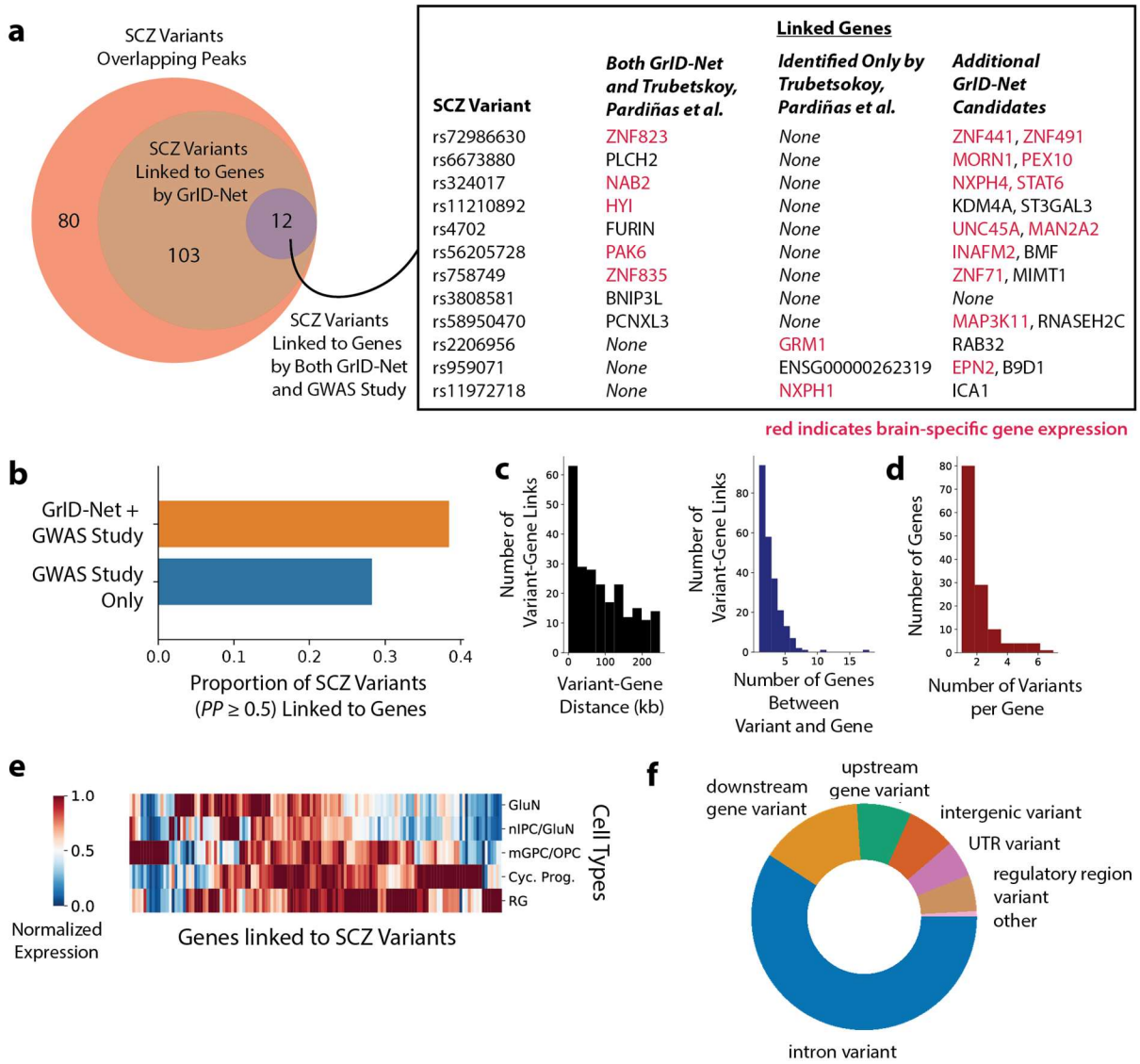


Figure 3. GrID-Net predictions link schizophrenia variants to affected genes. (a,b) Schizophrenia (SCZ) variants from Trubetskoy, Pardiñas *et al*., prioritized by GrID-Net. Overall, 195 SCZ variants overlapped with scATAC-seq peaks from the human corticogenesis dataset. Compared to the original study, GrID-Net enables the annotation of many additional variants in this subset. Furthermore, as multimodal single-cell assays become common, we expect GrID-Net's coverage to widen. We also compared specific SCZ variant-gene links for the 12 variants where the original study also reported an annotation. **(c)** GrID-Net is able to identify long-range variant gene associations: (Left) genomic distance between an SCZ variant and the

transcription start site (TSS) of its affected gene(s). (Right) number of genes in the genomic region between an SCZ variant and its affected gene(s). **(d)** Histogram of the number of variants linked to each SCZ gene. **(e)** Genes linked to SCZ variants are cell type-specific. We show the heatmap of normalized expression values for all genes linked to SCZ variants in cell types involved in human corticogenesis. **(f)** Distribution of annotated genomic locations for SCZ variants linked to genes. A substantial fraction of genes identified by GrID-Net are associated with intronic variants which have been reported to have tissue-specific regulatory effects.

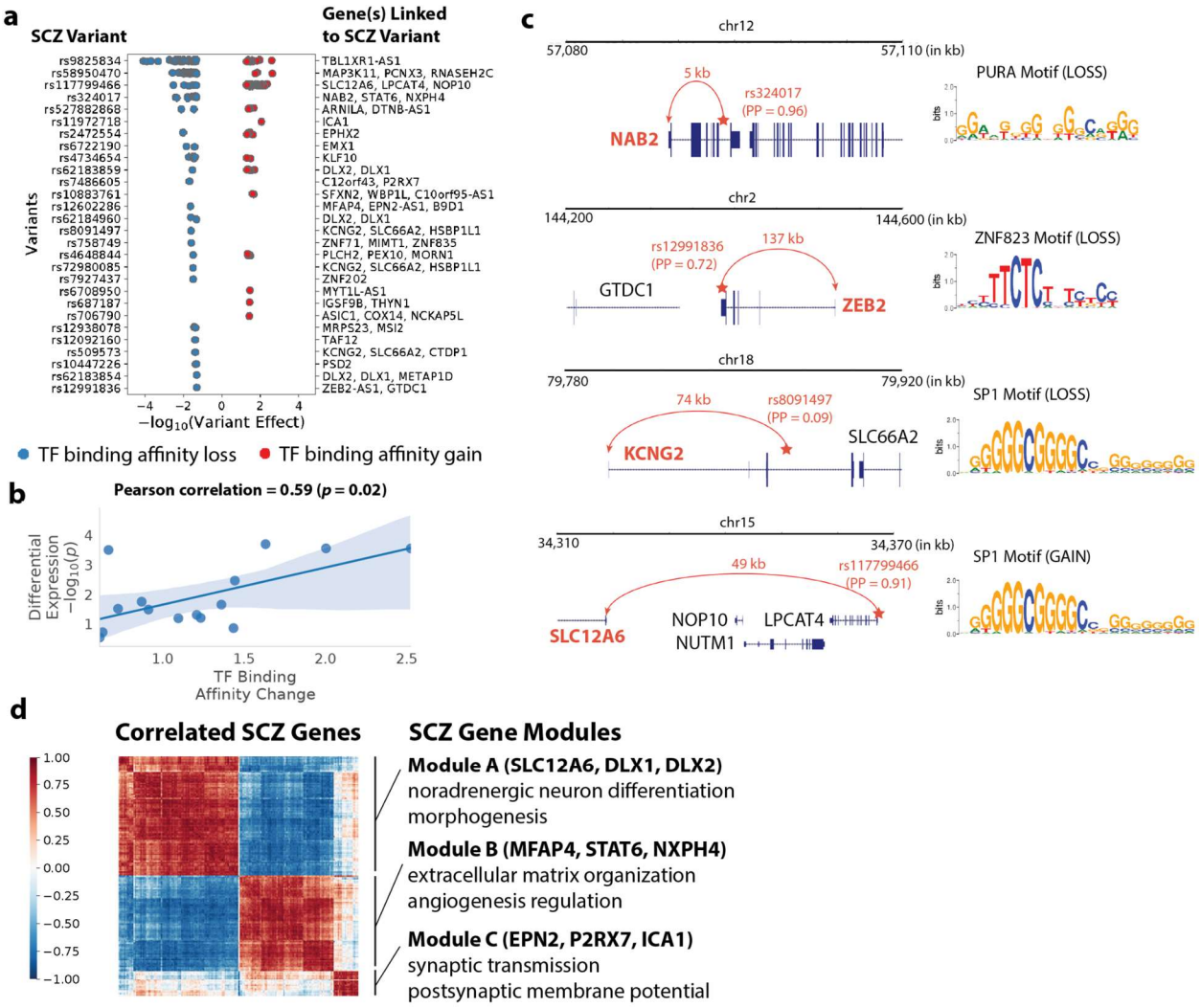


Figure 4. GrID-Net links transcription factor binding disruptions to gene dysregulation in schizophrenia. (a) TFs with at least a 20-fold change in predicted binding affinity caused by SCZ variants and the genes linked to these variants. Many of these TFs (e.g., KLF and SP families) are implicated in SCZ. **(b)** Relationship between TF binding affinity changes caused by SCZ variants and significance of differential expression in their linked genes between patient-derived SCZ glial progenitor cells and matching controls. **(c)** Examples of SCZ variant-gene links supported by at least 20-fold changes in predicted TF binding affinity. **(d)** Gene modules co-expressed with SCZ genes in SCZ GPCs and matching controls.

698

Supplementary Figures

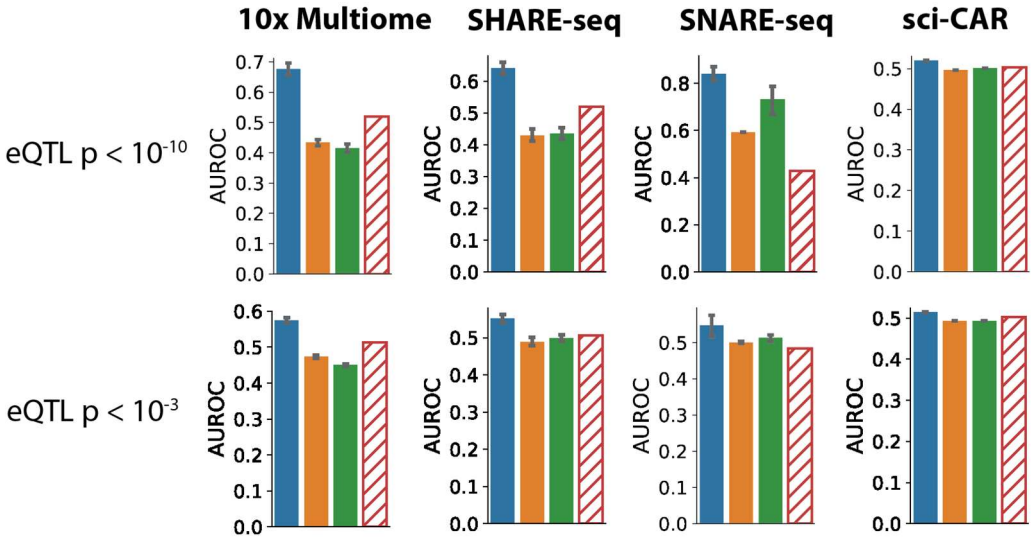


Figure S1. GrID-Net outperforms other methods in eQTL prediction accuracy across different eQTL significance thresholds.

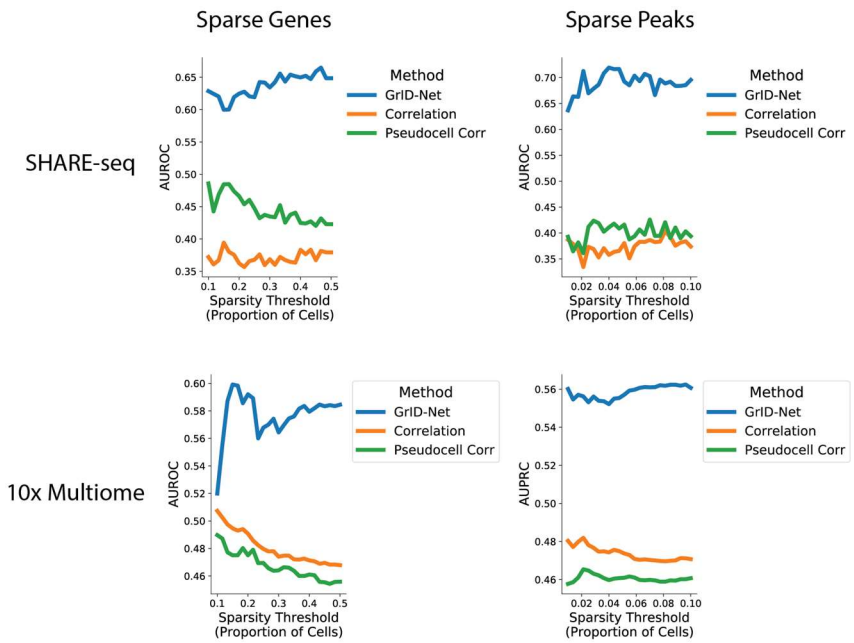


Figure S2. GrID-Net is robust to data sparsity in single-cell datasets.

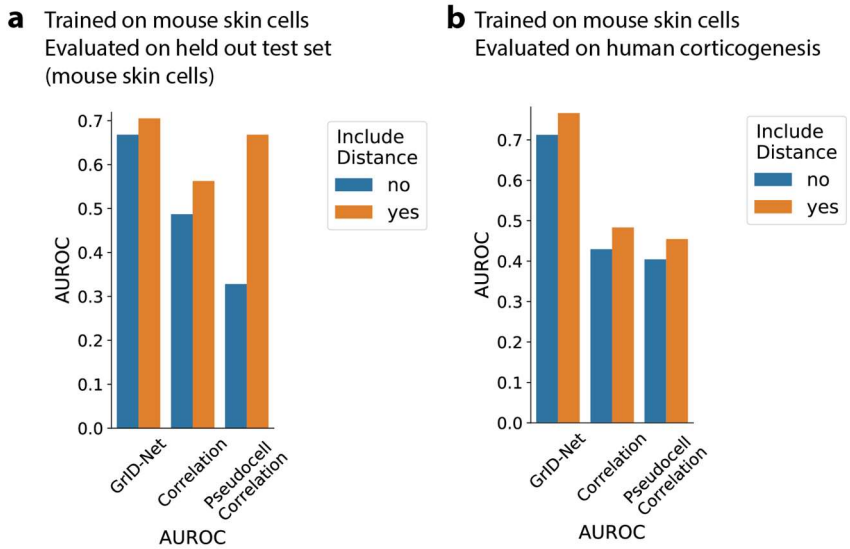


Figure S3. Genomic distance-dependent model improves peak-gene link prediction accuracy (a) within and (b) across species and cell types.

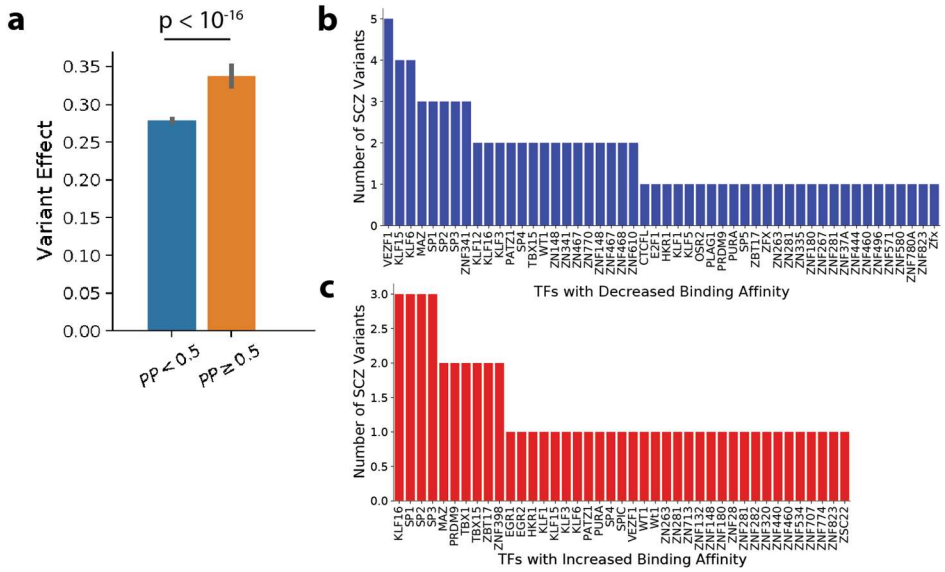


Figure S4. (a) Relationship between FINEMAP posterior probabilities (PP) for a SCZ variant and its corresponding TF binding affinity change (Welch's one-tailed *t*-test). TFs with the most widespread **(b)** increase or **(c)** decrease in binding affinity between the reference and the SCZ risk allele.