An econometric lens resolves cell-state parallax

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

[†] Contributed equally.

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

Abstract

Genome-wide association studies (GWAS) often report disease-linked genetic variation at noncoding genomic loci, but it is difficult to identify which genes these loci affect. Advances in single-cell multimodal assays that profile chromatin accessibility and gene expression in the same cell hold promise for addressing this challenge, as they can reveal causal locus-gene relationships. However, existing computational approaches for estimating such relationships from single-cell data are unable to account for what we refer to as "cell-state parallax," the time lag between epigenetic and transcriptional modalities due to their cause-and-effect relationship. In dynamic biological processes, chromatin regulatory potential is a manifestation of this parallax which we hypothesized could be recovered from single-cell snapshots. Leveraging the econometric concept of predictive temporal causality, we newly identify tissue-specific locus–gene associations where accessibility of the locus predicts expression of the gene. Our algorithm, GrID-Net, is a neural network-based generalization of classical Granger causal inference that enables graph-structured analysis, instead of being limited to sequential orderings. We applied it to a single-cell study of human corticogenesis to predict neuronal cis-regulatory elements, enabling us to interpret genetic variants in schizophrenia (SCZ). We identified 132 genes linked to common noncoding SCZ variants, including the potassium

[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

transporters KCNG2 and SLC12A6, and we present a strategy for unveiling the regulatory mechanisms underlying gene dysregulation. Our work points to the transformative potential of single-cell multimodal assays for discovering novel gene regulatory mechanisms and provides a general framework for linking genetic variants to gene dysregulation in disease.

Introduction

Dynamic changes in the chromatin states of cis-regulatory genomic elements are critical to the regulation of transcriptional programs across many biological processes, including development, cellular activation, reprogramming, and perturbation responses[1–3]. Correspondingly, genome-wide association studies (GWAS) have implicated thousands of noncoding loci in diseases and complex traits, pointing to their functional relevance[4,5]. However, uncovering mechanistic insights linking such noncoding loci to the genes they regulate remains a challenge. The common approach of simply associating a locus to its nearest gene neglects distal locus-gene interactions. Meanwhile, experimental approaches 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 approaches based on bulk data have also been proposed. For instance, Wang et al. integrated multiple data types[8], including chromatin contacts and disease-specific marker genes, to link variants to genes. Fulco et al.'s activity-by-contact (ABC) model[9,10] predicts the enhancers of a gene based on the activity of the enhancers and the 3D chromatin structure around the gene. However, none of these models are able to assess if the putative enhancer actually has any effect on its target gene's expression.

Recent advances in single-cell multimodal technology that can profile chromatin accessibility (scATAC-seq) and gene expression (scRNA-seq) in the same cell offer the opportunity to discover tissue-specific, genome-scale, causal relationships between noncoding loci and genes from a single experiment[11–14]. However, computational extraction of causal locus–gene associations from this data continues to be problematic. Current methods mainly rely on calculating contemporaneous correlations between scATAC-seq and scRNA-seq readouts in the same cell but this approach is limited as such correlations can not capture the causal asynchrony between the chromatin and gene expression modalities of the gene

60 regulatory hierarchy (**Figure 1a**). For instance, poised enhancers, which play essential roles in
61 development, display changes in chromatin states and histone modifications *prior* to induction of
62 their target genes[15]. Thus, the appropriate epigenetic and transcriptional readouts to be cross-
63 referenced come from slightly different cell states.

64 We refer to this asynchrony as "cell-state parallax," motivated by the optical concept of
65 parallax that describes the perceived displacement of an object when observed from two
66 different vantage points. In optics, having two similar-but-not-identical measurements of the
67 same entity can reveal new categories of insight. Consider cell phone cameras with dual lenses:
68 3D depth perception is unlocked by capturing 2D images that are nearly identical but have a
69 slight offset between them. Analogously in single-cell multimodal assays, we measure two (or
70 more) distinct modalities per cell that each capture a picture of the cell's state but at a slight
71 offset due to the lag between the epigenome and transcriptome. Combining these paired
72 readouts while accounting for their offsets can thus newly enable the detection of causal links
73 between the modalities. However, doing so is a challenge: cross-referencing the modalities
74 requires that local, per-cell measurements be resolved in an inference framework that can relate
75 measurements across cells to identify causal locus-gene associations while being robust to the
76 sparsity of both scATAC-seq (< 1% non-zero entries per cell) and scRNA-seq data[16,17]. To
77 address sparsity, some correlation-based approaches average the readouts of transcriptionally-
78 similar cells into pseudocells[11], but this comes at the cost of losing cell-level resolution and
79 heterogeneity, which are crucial in single-cell analyses.

80 To resolve single-cell snapshots of per-cell dynamics into genome-wide causal locus-
81 gene associations, we take inspiration from econometrics. There, a predictive relationship
82 between two time-varying processes, e.g. the effect of interest rates on inflation, can be inferred
83 solely from observational data analyzed across time. In principle, multimodal data synthesis
84 followed by trajectory and pseudotime analysis of a single-cell dataset can provide a cell-state
85 transition axis along which a similar predictive analysis could be applied to infer locus–gene

relationships. However, our biological setting presents a fundamental challenge. Unlike in econometrics, where observational data is ordered along a time series, single-cell measurements are not sequentially ordered, and if a trajectory has multiple branches, the pseudotime axes across branches are not comparable. As a result, classical econometric techniques like Granger causal analysis can not be used in their standard formulation. Our key conceptual advances are two complementary innovations that unlock such predictive analyses: a) a directed acyclic graph (DAG)-based representation of single-cell data that encodes *local* cell-state transitions, and b) a machine learning extension of classical Granger causality that can operate under the partial ordering assumptions supported by the DAG.

Here, we present GrID-Net (Granger Inference on DAGs), a novel framework for causal inference on DAGs that we apply to single-cell multimodal data to infer regulatory locus-gene relationships (**Figure 1b,c**). Leveraging lagged message-passing on a graph neural network, GrID-Net generalizes Granger causality from sequential orderings to DAG-structured orderings. Unlike current approaches, it can detect asynchronous and nonlinear interactions since it is able to account for the graph-structured dynamics of single-cell data. We demonstrate the improved accuracy of noncoding locus-gene links inferred from GrID-Net compared to existing approaches across cell types and single-cell multimodal assay methods. We then apply GrID-Net to interpret fine-mapped genetic variants associated with schizophrenia (SCZ), associating these variants with target genes. We predict key gene mechanisms and modules underlying SCZ disease pathogenesis (**Figure 1d**). Our findings demonstrate the utility of GrID-Net as a powerful tool for uncovering the causal mechanisms of disease and development. Our work also illustrates the power of multimodal single-cell studies for investigating the impact of genetic variants on function and phenotype. The code for GrID-Net can be found at https://github.com/alexw16/gridnet and the full list of SCZ variant-gene links can be found in **Supplemental Table 1**.

Overview of GrID-Net

In general, inferring true causality can be a challenge as it requires perturbations to confirm cause-and-effect relationships and validations to rule out potential confounders. Granger causality is a powerful econometric concept that can *approximate* causality in time-varying processes solely from observational data. It leans on the principle that a cause should occur before its effect[18,19]. Given two time-dependent variables x (potential cause) and y (effect), if the forecast for y can be improved by accounting for the past values of x, then x is said to Granger cause y since changes in x temporally precede those in y.

One of our key contributions is to broaden the set of domains to which Granger causal analyses can be applied. Classical Granger causal inference requires a set of observations to be arranged in a total, or sequential, ordering, such as in time series data. We generalize Granger causal inference to settings where only a partial ordering is available (i.e. the set of events in the dynamical system can be represented as a directed acyclic graph (DAG)), with branches and merges but no cycles. Our graph neural network method for Granger causal inference, GrID-Net, can be applied to any DAG-structured dynamical system, such as those represented by single-cell trajectories, social media relationships, and financial transaction networks.

To model the asynchrony between cause and effect, we modify standard graph neural networks to allow for lagged message passing. This modified architecture accumulates information about past states, leveraging the connectivity of these past states within a graph to forecast the present state in a manner that is aware of the graph-structured dynamics. GrID-Net consists of two forecasting models: a *reduced* model forecasts an effect y using only the past of y and a *full* model forecasts the effect y using both the past of y and the past of a candidate Granger causal variable x. The strength of the Granger causal link between x and y can be

evaluated by the statistical F-test which quantifies the improvement in forecasting accuracy of the *full* model over that of the *reduced* model.

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

GrID-Net more accurately identifies noncoding locus-gene links

We evaluated GrID-Net on four single-cell multimodal datasets (scATAC-seq and scRNA-seq) that characterize a range of biological processes, including cell differentiation and drug treatment response[11–14]. For each dataset, we sought to validate against expression quantitative trait (eQTL) loci profiled in matching tissue (**Methods**). As eQTLs represent associations between genetic variation at a locus and the corresponding changes in a gene's expression, they serve as useful genome-wide proxies for causal interactions between genomic loci and genes.

For each gene, we intersected all ATAC-seq peaks within 1 Mb of its transcription start site (TSS) with the set of eQTL variants linked to it, associating each peak with the most significant variant overlapping it. We then evaluated the peak–gene pairs prioritized by GrID-Net and two correlation-based methods used in previous work[11,21], Pearson correlation and pseudocell correlation. The latter computes correlations on measurements aggregated across

161 multiple cells. We also evaluated the ABC model's predictions on bulk data from matching
162 tissues, comparing it against the predictions from the single-cell datasets (**Methods**).
163 We evaluated the accuracy of each method in retrieving the peak–gene pairs that
164 overlap significant eQTLs ($P < 10^{-10}$); our findings were robust to the choice of the eQTL
165 significance threshold (**Figure S1**). GrID-Net was significantly more accurate than the other
166 methods across all datasets (**Figure 2a**), exhibiting 7-27% higher area under the receiver
167 operating characteristic curve (AUROC) in 3 of the 4 datasets. GrID-Net demonstrated the
168 smallest improvement in AUROC for the sci-CAR dataset, although all methods performed
169 poorly on this dataset, likely due to the lower data quality of sci-CAR as the earliest single-cell
170 multimodal assay developed. Meanwhile, the correlation-based methods exhibited near-random
171 performance across all datasets, consistently underperforming not only GrID-Net but also the
172 bulk data-based ABC model. The low accuracy of predictions by correlation-based methods,
173 contrasted with the substantial improvement offered by GrID-Net, demonstrates the potential to
174 unravel much richer regulatory insights from these multimodal assays than has been possible.
175 We next assessed the robustness of the single-cell methods to sparsity in single-cell
176 data. GrID-Net better handled sparse data (**Figure S2**), yielding higher accuracies than the
177 correlation-based methods for peak–gene predictions involving genes with sparse data (non-
178 zero transcripts in fewer than 10% of cells) or sparse peaks (under 2% of cells). Notably,
179 pseudocell correlation failed to consistently outperform standard Pearson correlation on these
180 sparse peak–gene cases, even though the former is explicitly designed to compensate for the
181 sparsity of measurements at the level of individual cells. This may be because pseudocell
182 averaging reduces the effective number of independent observations and thus relinquishes the
183 statistical advantages of single-cell heterogeneity. Altogether, these analyses indicate not only
184 GrID-Net's robustness to the current sparsity of single-cell data but also the potential for its
185 increased effectiveness as the sensitivity of these multimodal assays continues to improve.

186 To further benchmark GrID-Net using functional indicators of noncoding locus–gene
187 interactions, we examined three-dimensional chromatin contacts (from bulk Hi-C data) between
188 gene promoters and noncoding loci. Compared to the two correlation-based methods, we found
189 that GrID-Net more accurately prioritized peak–gene pairs that were in frequent three-
190 dimensional contact (**Figure 2b)**. The ABC method uses bulk Hi-C data itself as an input so we
191 did not evaluate its agreement with chromatin contacts. In future, the GrID-Net framework could
192 be expanded to also utilize Hi-C data when available— as we describe later, incorporating one-
193 dimensional sequence proximity as a signal into GrID-Net improves its results. However, Hi-C
194 data is currently available only for a limited set of tissues[22] and the ability to make regulatory
195 inferences from just one multimodal assay enables the convenient study of rare cell types,
196 cancer mutations and drug responses etc.
197
198 **GrID-Net reveals how chromatin accessibility primes gene activity**
199 To better understand the biological basis of GrID-Net's superior accuracy, we applied it on a
200 single-cell multimodal dataset profiling human corticogenesis. We then honed in on peak–gene
201 pairs where GrID-Net and correlation-based methods disagree. Specifically, peak–gene pairs
202 were ranked according to the difference between their GrID-Net and correlation-based
203 prioritization scores (**Methods**), with higher ranking pairs being the ones that were differentially
204 prioritized by GrID-Net. We then investigated the distinguishing biological characteristics of
205 these differentially-prioritized pairs.
206 One of the motivations behind our design of GrID-Net is the observation that regions of
207 chromatin around lineage-determining genes become accessible prior to gene expression,
208 foreshadowing lineage choice[23,24]. GrID-Net was designed to capture this regulatory potential of
209 accessible chromatin, and we assessed if it is able to do so. We first associated each peak and
210 each gene with the pseudotime corresponding to its highest value to determine the lag
211 associated with each peak-gene pair. For both correlation-based methods, we found that peak–

212 gene pairs differentially prioritized by GrID-Net are marked by significantly longer pseudotime
213 lags between chromatin accessibility and gene expression ($P < 10^{-26}$, Welch's one-tailed *t*-test,
214 **Figure 2c**, **Methods**). This result confirms GrID-Net's ability to account for the asynchrony
215 between noncoding regulatory activity and downstream gene expression control that are
216 expected to characterize causal interactions between the two. To explore the implications of
217 GrID-Net's ability to account for this asynchrony, we obtained data on histone modification
218 states of enhancers in neural progenitor cells and cross-referenced it with ATAC-seq peaks[25]. In
219 peak–gene pairs differentially prioritized by GrID-Net, more poised enhancers (H3K4me1 only)
220 overlapped with the peaks than active enhancers (both H3K4me1 and H3K27ac)[26] (**Figure 2d**,
221 **Methods**). This was not the case for peak–gene pairs differentially prioritized by correlation-
222 based methods. Poised enhancers are regulatory sites that can trigger future gene expression
223 changes but have not yet affected expression[26,27]; in contrast, active enhancers correspond to
224 ongoing transcription[28]. This suggests that GrID-Net is indeed capturing chromatin regulatory
225 potential and — towards understanding the causal mechanism of regulation — is better able to
226 detect the lagged relationships between the activities of enhancers and genes.
227 To evaluate the biological significance of capturing these likely-causal interactions, we
228 examined the functions of genes implicated in peak–gene pairs differentially prioritized by GrID-
229 Net. We found these genes to be highly enriched for gene sets associated with neurogenesis
230 regulation and neural progenitor cells ($FDR < 10^{-7}$, GSEA Preranked[29]), directly relating to the
231 process of corticogenesis profiled in this dataset. We emphasize that the enrichment score is
232 calculated only on the peak–gene pairs that GrID-Net prioritizes over the correlation-based
233 methods. In comparison, the set of genes from peak–gene pairs that *all* methods score highly is
234 functionally enriched for housekeeping genes rather than neuronal activity, suggesting that
235 GrID-Net's ability to detect poised enhancers is key to discovering cell-type specific regulatory
236 relationships. Gene expression control during differentiation has been reported to depend highly
237 on asynchronous enhancer-gene interactions[30]. The enrichment of these differentiation-related

238 gene sets thus underscores GrID-Net's ability to construct accurate genome-wide maps of
239 noncoding locus–gene links that go beyond correlative analyses and towards causality.
240
241 **Linking variants to schizophrenia disease genes**
242 The ability to link noncoding loci to the genes they regulate enables the discovery of molecular
243 mechanisms underlying complex diseases. Genome-wide association studies (GWAS) have
244 identified thousands of loci associated with diseases[31]. However, over 90% of GWAS variants
245 lie in noncoding regions[32], and interpreting the functions of these noncoding variants involves
246 the challenge of linking them to the specific genes that they regulate in a cell type-specific
247 manner. With GrID-Net, we can leverage data from a single-cell assay on matching tissue to
248 uncover these cell type-specific variant-gene links and help bridge the gap between genetic
249 variants and gene dysregulation in disease.
250 We used GrID-Net to study variants associated with schizophrenia (SCZ), a psychiatric
251 disorder for which inherited genetic variants are believed to underlie its pathogenesis[33]. SCZ's
252 high prevalence of around 1%, its substantial morbidity and mortality, and the lack of effective
253 treatments are largely due to our limited understanding of SCZ pathophysiology[34,35]. We applied
254 the GrID-Net peak-gene links inferred from human corticogenesis[21], to interpret and analyze
255 variants reported in the largest GWAS of SCZ to date[36]. We refer to this original study as the
256 Trubetskoy, Pardiñas *et al*. study.
257 To boost GrID-Net's accuracy on shorter-range peak–gene interactions, we incorporated
258 one-dimensional genomic distance as an additional feature for predicting peak–gene
259 associations. We observed a higher likelihood of peak–gene interaction if the two are located
260 close to each other, in line with reported patterns of genomic distance-dependent regulatory
261 effects[37,38]. We calibrated the relative importance of this feature on a dataset profiling mouse
262 skin differentiation[24] where we fitted a logistic regression model that combines the raw GrID-Net
263 score with a genomic distance score (**Methods**). We found that integrating genomic distance

264 with GrID-Net improved predictions of peak–gene links for mouse skin cells. The discriminative
265 power on short-range peak–gene interactions improved, while the ability to detect long-range
266 interactions was preserved. Furthermore, this model trained on mouse skin cell data performed
267 well across species and cell types in human corticogenesis as well (**Figure S3**). This genomic
268 distance-dependent model's scores were used for all remaining analyses of the human
269 corticogenesis dataset.

270 From the SCZ GWAS study, we retrieved all fine-mapped genetic variants with posterior
271 probability (*PP*) over 5%. We intersected these variants with peaks from the human
272 corticogenesis dataset. We then generated an initial set of candidate SCZ genes by associating
273 each variant with up to 3 of its overlapping peak's highest scoring genes, retaining only variant–
274 gene associations with scores in the top 50% of all peak–gene scores. This approach yielded a
275 set of 132 genes that not only is enriched for genes implicated in the original SCZ GWAS study
276 (*FDR* = 0.037, Enrichr[39]) but also includes novel SCZ gene candidates (**Supplemental Table**
277 **1**). For example, of the 12 SCZ variants that were prioritized by both GrID-Net and Trubetskoy,
278 Pardiñas *et al*., 9 were linked to the same target genes by both approaches (**Figure 3a)**. This
279 overlap included the association between rs324017 and NAB2, a gene implicated in stress-
280 induced neuroprotective maintenance that also is the downstream target of the SCZ drug
281 cinnarizine[40,41]. For these same 12 SCZ variants, GrID-Net also proposed 20 additional SCZ
282 candidate genes, of which 12 (60%) were found to be specifically expressed in brain tissues and
283 thus potentially relevant in SCZ etiology (**Methods**). Some of these additional candidates
284 include NXPH4, a neurexophilin involved in synapse function and cerebellar motor control[42],
285 and EPN2, an epsin linked to neural fate commitment[43]. Moreover, of the 78 SCZ variants with
286 *PP* over 50%, only 22 (28%) were linked to genes by the original SCZ GWAS study (**Figure 3b**).
287 Inclusion of SCZ variants that were linked to genes by GrID-Net increased this total to 30 (38%)
288 SCZ variants. GrID-Net can, therefore, be used to complement standard strategies for linking
289 variants to genes, such as those used in the original SCZ GWAS study. It can propose gene

290 associations for variants with no annotations and also augment existing annotations by
291 suggesting additional genes that a variant may be affecting.
292 In particular, GrID-Net helps address a key challenge in GWAS analyses: the common
293 strategy of assigning GWAS variants to their closest gene[44] misses many biologically
294 meaningful variant–gene relationships. By leveraging single-cell heterogeneity, GrID-Net can
295 evaluate longer range interactions. Many of the variant–gene links we identify are characterized
296 by large genomic distances between the variant and the transcription start site (TSS) of the
297 linked gene. Moreover, the variants are often linked to genes that are multiple genes over from
298 their closest gene (**Figure 3c**). For example, DCP1B and NCKAP5L, each of which are
299 connected to only a single variant, are separated from their respective variants by 18 and 11
300 genes (235 and 242 kilobases (kb)), respectively. DCP1B has been implicated in SCZ,
301 hypomania, and memory performance[45,46], while NCKAP5L, a cytoskeletal gene involved in
302 neuronal activity[47], has previously been associated with SCZ as well as amyotrophic lateral
303 sclerosis[48], indicating possible roles for these genes in neuron-specific dysregulation in SCZ.
304 In addition, 52 candidate genes (39.4% of the overall set) were associated with more
305 than one variant (**Figure 3d**). These genes may be controlled by the coordinated activity of
306 multiple regulatory elements and co-regulators relevant to SCZ[49]. IGSF9B was linked to 7
307 noncoding SCZ variants – the most of any gene. Coding variants in IGSF9B have been found to
308 be associated with major depression and the negative symptoms of SCZ[50,51], underscoring its
309 role in regulating affective symptoms in SCZ. Similarly, KCNG2, a potassium channel-encoding
310 gene, was linked to 6 noncoding SCZ variants, and has been proposed as a drug target for SCZ
311 given the known disruption of KCN-related pathways in SCZ[35,52,53]. Neither of these genes were
312 prioritized in the original GWAS analysis.
313 We also assessed the cell-type specificity of the genes linked to SCZ variants. We found
314 that 104 candidate SCZ genes (78.8% of the overall set) are differentially expressed across the
315 cell types profiled in the human corticogenesis dataset (Welch's *t*-test with Benjamini-Hochberg

316 correction $P < 0.05$, **Figure 3e**). Further analysis of the SCZ variants also revealed that 59% of
317 the variants lie in intronic regions, while only 7% of variants lie in intergenic regions more than 5
318 kb from any gene (**Figure 3f**, **Methods**). As enhancers with tissue-specific effects are highly
319 enriched in introns while intergenic enhancers more often regulate housekeeping genes[54], these
320 findings support the potential cell type-specificity of the regulatory disruptions underlying SCZ
321 pathogenesis.
322
323 **Transcription factor binding disruptions are implicated in schizophrenia**
324 Transcription factor (TF) binding at the cis-regulatory elements located in noncoding genomic
325 loci is a key mechanism of gene regulation. We observed that SCZ variants with high *PP* are
326 associated with larger TF binding affinity changes (**Figure S4a**). To investigate the
327 dysregulatory role of SCZ variants prioritized by GrID-Net, we therefore evaluated their effect on
328 TF binding. For each variant associated with a gene by GrID-Net, we calculated the difference
329 in the predicted TF binding affinity between the reference allele and the SCZ risk allele
330 (**Methods**, **Supplemental Table 2**). Of the 115 such SCZ variants, 28 (these were associated
331 with 49 target genes) were predicted to cause a 20-fold or stronger change in at least one TF's
332 binding affinity (**Figure 4a**). The TFs we identified as having the most widespread binding
333 affinity changes are known to be implicated in SCZ, exemplifying GrID-Net's ability to deliver de
334 novo mechanistic insight. VEZF1, a vasculature-related TF reported to play a role in SCZ
335 dysregulation[55], displayed diminished binding affinity most frequently (in 5 variants). In addition,
336 KLF15 and KLF6, two Kruppel-like family (KLF) TFs known to regulate neuronal growth and
337 linked to SCZ[56,36], were each associated with reduced binding affinity in 4 variants (**Figure S4b**).
338 Meanwhile, the TFs with the most widespread increase in binding affinity – SP1, SP3, and
339 KLF16 – have also been reported to be disrupted in SCZ[57,58] (**Figure S4c**). More broadly, of the
340 TFs with a 20-fold binding affinity change in at least one variant, 57.9% (=33/57) have cell type-
341 specific expression patterns in corticogenesis (Welch's *t*-test with Benjamini-Hochberg

342 correction $P < 0.05$). Collectively, these results point to genetic disruption of neuron-specific
343 regulatory elements as potential drivers of SCZ pathogenesis.
344 We reasoned that if the TF binding affinity changes at a SCZ variant are phenotypically
345 meaningful, they should result in expression changes of the corresponding gene(s) (as identified
346 by GrID-Net). Accordingly, we obtained gene expression data from glial progenitor cells (GPCs)
347 produced from SCZ patient-derived induced pluripotent cells, as well as matching controls[53,59],
348 and computed differential expression patterns in SCZ GPCs. Focusing on glial differentiation
349 genes, we observed that the differential expression of a gene in SCZ GPCs was indeed
350 significantly correlated with TF binding affinity changes at the corresponding SCZ variant
351 (Pearson correlation coefficient = 0.59, **Figure 4b**, **Methods**). The consistency between these
352 independent observations provides support to our overall analytic approach and underscores
353 the potential to leverage GrID-Net for examining causal, gene-specific regulatory mechanisms.
354 We drilled down into specific genes reported by our analysis, focusing on those linked to
355 SCZ variants associated with at least 20-fold TF binding affinity changes (**Figure 4c**). NAB2, the
356 aforementioned gene implicated in neuroprotective processes[41], was linked to rs324017, which
357 contributes to a 200-fold decline in binding affinity for the TF PURA. Intriguingly, PURA is also
358 critical for neural function and development, and its dysfunction is implicated in several
359 neurological and cognitive impairments[60,61]. In addition, rs12991836 was linked to ZEB2, whose
360 TSS is positioned 137 kb from the variant. ZEB2 has been previously reported as a target of
361 long-range regulation, and heterozygous ZEB2 mutations are known to cause Mowat-Wilson
362 syndrome, which features structural brain abnormalities and intellectual disability[62].
363 Furthermore, GrID-Net prioritized two potassium transport genes - KCNG2 and
364 SLC12A6 - a class of genes which has been associated with compromised potassium uptake
365 when dysregulated[53]. This phenotype is characteristic of not only SCZ but also other neuron-
366 related diseases, such as Huntington's disease, multiple sclerosis, and hemiplegic migraine[59,63].
367 KCNG2 was linked to 6 SCZ variants, the second most of any gene as described earlier. Using

368 TF binding affinity changes to hone in on the most relevant variants, we identified rs8091497 as
369 the variant with the largest TF binding affinity change. rs8091497 was associated with a 40-fold
370 decline in binding affinity for SP1, which itself has been linked to behavioral abnormalities in
371 SCZ[58]. As for SLC12A6, its corresponding variant (rs117799466) increases the binding affinity
372 for many TFs with repressive capabilities. The three TFs with the greatest binding affinity gains
373 at this locus – SP1, ZN281, and MAZ – all play a role in neuronal differentiation[58,64,65], and both
374 SP1 and ZN281 have also shown increased expression levels in SCZ as well[66,67]. Combined
375 with their increased binding affinity incurred by rs117799466, the upregulation of these TFs may
376 contribute to the repression of SLC12A6, an outcome that has been previously linked to the
377 potassium uptake impairments that characterize SCZ glia[53]. Together, these findings suggest a
378 possible mechanism for SLC12A6 downregulation in SCZ that implicates enhanced binding of
379 neuron-related repressors to a SLC12A6-targeting enhancer to reduce its expression levels.
380 We then broadened our analysis to gene modules, exploring if the genes identified by
381 GrID-Net are systematically involved in SCZ-related pathways. The 14 SCZ variants that cause
382 at least 40-fold changes in TF binding affinity are linked by GrID-Net to 26 genes, whose
383 differential expression patterns we examined in SCZ GPCs. We detected 3 distinct gene
384 modules that are co-expressed with several of these genes (**Figure 4d**, **Methods**). Module A
385 (SLC12A6, DLX1, and DLX2) is enriched for genes related to noradrenergic neuron
386 differentiation (*FDR* < 0.05) and morphogenesis (*FDR* < 0.02); Module B (MFAP4, STAT6,
387 NXPH4) is enriched for extracellular matrix organization (*FDR* < 10^{-14}) and angiogenesis
388 regulation (*FDR* < 0.02) genes, and Module C (EPN2, P2RX7, ICA1) is enriched for synaptic
389 transmission (*FDR* < 10^{-4}) and postsynaptic membrane potential (*FDR* < 0.05) (Enrichr[39]). Each
390 of these pathways has been previously associated with SCZ. Noradrenergic neurons have long
391 been implicated in disrupted neuromodulation in SCZ[68], neuronal angiogenesis underlies the
392 vascular abnormalities observed in SCZ[69], and genetic studies have pointed to the critical role of
393 synaptic function in SCZ[36]. These pathways, therefore, enable the interpretation of SCZ variants

394 beyond the dysregulation of the specific genes that they directly affect, presenting an
395 opportunity to discover systems-level mechanisms of SCZ pathogenesis.
396
397 **Discussion**
398 Most methods for single-cell multimodal data integration have focused on harmonization across
399 modalities to improve, for instance, cell type identification. With this work, we demonstrate it is
400 also powerful to systematically study the deviations between the modalities. We presented
401 GrID-Net, a method to uncover causal connections between genes and noncoding genomic loci
402 using co-assayed single-cell RNA-seq and ATAC-seq data. GrID-Net operates at a genome-
403 wide scale, is tissue-specific, and can effectively detect long-range (up to 1 Mb) cis-regulatory
404 associations. We show how it can supplement existing approaches for interpreting GWAS
405 results and prioritizing disease-linked genes. There, a longstanding challenge has been
406 identifying causal variant-gene associations specific to the tissue of interest. With GrID-Net, a
407 single observational assay of the matching tissue can power a genome-wide estimation of
408 variant-gene relationships. We demonstrate this by leveraging an observational single-cell
409 assay of human corticogenesis to interpret the functional effects of SCZ variants. The use of
410 GrID-Net enhances both the breadth and depth of gene prioritization analysis: we were able to
411 find associations for many variants that were previously unlinked while also augmenting existing
412 annotations. For instance, we identified SCZ variants causing large (>20-fold) changes in TF
413 binding affinity and connected them to 49 genes. Many of these genes are located distally from
414 the SCZ variants that are predicted to affect them and would have been neglected by existing
415 GWAS variant–gene associations analyses that consider only genomic proximity. Furthermore,
416 by correlating TF binding affinity changes with differential expression in SCZ glial progenitor
417 cells, we predicted possible mechanisms of TF-driven dysregulation for genes like KCNG2 and
418 SLC12A6. Combined with an exploration of SCZ-related pathways associated with GrID-Net-
419 identified genes, these analyses provide a comprehensive strategy for unveiling mechanistic

420 insights about SCZ pathogenesis. Such analyses can also be applied to investigate other
421 diseases.
422 Our work rests on two conceptual insights that we believe are broadly applicable. First,
423 when different modalities are measured in the same cell their causal relationship implies a
424 temporal offset between them, a phenomenon we refer to as “cell-state parallax”. Here, we
425 exploit the parallax between RNA-seq and ATAC-seq, but such parallax can also be leveraged
426 from other co-assayed modalities such as chromatin structure or methylation. We view RNA
427 velocity[70] and protein velocity[71] as specific applications of the broader concept of cell-state
428 parallax, the distinction being that velocity uses pre-defined mappings between modalities (e.g.,
429 spliced vs. unspliced transcript-counts for a particular gene) to infer the differentiation direction
430 of each cell. In contrast, such mappings between modalities may be unclear in general cell-state
431 parallax and require customized inference approaches. Accordingly, our second conceptual
432 insight is to develop computational frameworks that can accommodate local measures of cell
433 differentiation. Much of the existing work on modeling gene expression dynamics utilizes
434 mathematical and statistical tools designed for time-series analysis[72,73]. While these methods
435 can be effective in certain settings, they make the unrealistic assumption that a globally-valid
436 sequential ordering of cells is always feasible. In contrast, we characterize dynamics locally
437 around each cell, formulating the differentiation landscape as a directed acyclic graph capable
438 of handling forks and merges in trajectories. We expect a promising direction of future research
439 will be computational inference techniques, like GrID-Net, that can operate under this more
440 realistic assumption.
441
442 Since GrID-Net operates on observational data, the locus-gene associations it infers are
443 causal hypotheses that need to be validated by perturbations. In the eQTL-based evaluation, we
444 sought to mimic such validations by investigating random genetic variants observed in the
445 population and cross-referencing them to observed expression changes. We also note that such

446 causal interactions are likely indirect and mediated by additional biophysical events, such as
447 transcription factor binding. For studying a particular disease, GrID-Net should be paired with a
448 single-cell assay of a matching tissue undergoing a dynamic biological process, such as
449 development, differentiation, disease, or drug response. GrID-Net can then translate the results
450 of this single assay to genome-wide causal hypotheses specific to the disease.

451 With the growing availability of single-cell multimodal datasets, our approach is
452 positioned to serve as a key framework for unlocking new mechanisms of gene regulatory
453 dynamics that underlie fundamental biology and disease. Moreover, GrID-Net can be applied
454 broadly to other domains featuring graph-based dynamics, e.g. citation networks to learn the
455 key drivers of a paper's impact. A key technical advance of this work is generalizing the
456 applicability of standard Granger causal analysis to a wider set of domains than previously
457 possible. We achieve this by leveraging graph-theoretic deep learning techniques to implement
458 the core Granger causal concept of lagged prediction on any DAG-structured ordering of
459 observations. Our work, like others before us[74], also highlights how modern deep learning
460 techniques can augment — rather than supplant — classical statistical formulations.
461
462
463
464
465
466

467 **Methods**

468

469 **GrID-Net formulation**

470 See Supplementary Note.

471

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

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

479

480 **Single-cell multimodal datasets**

481 We used four single-cell multimodal datasets that spanned a range of different multimodal
482 profiling techniques: 1) a 10x Multiome dataset profiling human corticogenesis, 2) a SHARE-seq
483 dataset profiling mouse skin differentiation, 3) a SNARE-seq dataset profiling mouse neonatal
484 cerebral cortical cells, 4) a sci-CAR dataset profiling A549 cells treated with dexamethasone.
485 For each of these datasets, we followed recommended practice[75] and applied $log\ (1 +$
486 $CPM/100)$ and $log\ (1 + CPM/10)$ transformations to the RNA-seq and ATAC-seq counts data,
487 respectively, where CPM indicates counts per million. We then filtered out all genes that were
488 represented in fewer than 1% of cells in the dataset and all peaks that were represented in
489 fewer than 0.1% of cells. We assembled a set of candidate peak-gene pairs by selecting all
490 peaks with centers within 1 Mb from each gene's TSS. For the 10x Multiome dataset, we only
491 considered peak-gene pairs that were separated by 250 kb to cohere with the genomic distance

threshold used in the original study[21]. To consider multiple samples for each dataset, we generated 5 uniformly sampled sets of 5,000 cells for each of the 10x Multiome, SHARE-seq, and SNARE-seq datasets, while we generated 5 uniformly sampled sets of 3,000 cells for the sci-CAR dataset due its smaller quantity of cells (3,081 cells).

Directed acyclic graph (DAG) construction

For each sample of a dataset, we first generated per-cell representations for each of the gene expression and chromatin accessibility modalities. Gene expression representations were constructed by (1) scaling the transformed RNA-seq data to unit variance and zero mean, (2) retaining the top 2000 most highly variable genes, (3) and projecting these normalized expression values onto their top 100 principal components (excluding components that have Spearman correlation $\rho > 0.9$ with the total counts per cell. Chromatin accessibility representations were obtained transforming the ATAC-seq data using tf-idf[76,77] and projecting the resulting values onto the top principal components as in (3) above. We then used Schema[20] to learn cell representations that unify information from these gene expression and chromatin accessibility representations. We constructed a k-nearest neighbor graph ($k = 15$) based on the Euclidean distances between these cell representations and inferred pseudotime using the scanpy[78] implementation of diffusion pseudotime[79] with default parameters. We then removed edges in the k-nearest neighbor graph if they did not align with the direction of increasing pseudotime, thus ensuring that no cycles exist and the resulting graph is a DAG.

Alternative methods

Pearson correlation. We computed the Pearson correlation between the transformed ATAC-seq and RNA-seq values for each candidate peak-gene pair, and the absolute value of the correlation coefficients were used to score the peak-gene pairs.

518 *Pseudocell correlation*. For each sample, we identified the 49 nearest neighbors for each cell
519 based on the Euclidean distances between the joint representations learned from Schema (see
520 "Directed acyclic graph construction"). We constructed a "pseudocell" from each cell and its 49
521 nearest neighbors, constructing its RNA-seq and ATAC-seq profiles by summing the raw counts
522 from each of the modalities for all cells included in the pseudocell. We then applied $log\ (1 +$
523 $CPM/100)$ and $log\ (1 + CPM/10)$ transformations to the RNA-seq and ATAC-seq counts data,
524 respectively. Pearson correlation was applied to the transformed ATAC-seq and RNA-seq
525 values for each candidate peak-gene pair, and the absolute value of the correlation coefficients
526 were used to score the peak-gene pairs.
527
528 *Activity-by-contact (ABC) model*. We obtained ABC prediction scores in cell types matching the
529 cell types profiled in the single-cell multimodal studies. For each single-cell study, we
530 intersected the candidate peak-gene pairs with the ABC enhancer-gene pairs, assigning each
531 peak-gene pair the corresponding ABC score. If no ABC enhancer-gene pair overlap was
532 identified, the peak-gene pair was assigned a value of 0, as the ABC prediction scores were
533 available only for enhancer-gene pairs that exceeded the threshold of 0.015 defined in the
534 original study[10]. For single-cell studies that profiled mouse cells, we mapped ABC enhancer-
535 gene pairs in matching human cell types to the mouse genome due to the strong conservation
536 of epigenomes between human and mouse[80,81].
537
538 **eQTL evaluations**
539 We obtained eQTL data from the eQTL Catalogue[82] and used eQTL data specific to the cell
540 type profiled in each single-cell multimodal study. For mouse tissues, we mapped eQTL results
541 for matching human tissues to the mouse genome, as human and mouse epigenomes show
542 strong conservation[80,81].

We first intersected eQTL variant-gene pairs with peak-gene pairs from the single-cell data and assigned each remaining peak-gene pair the most statistically significant eQTL. We then binned these peak-gene pairs according to the genomic distance between each gene's TSS and each peak's center, creating a bin for each of the following ranges of genomic distances: [1 bp, 1 kb), [1 kb, 10kb), [10kb, 100kb), [100kb, 1 Mb). We identified all peak-gene pairs that overlapped significant eQTL variant-gene pairs ($P < 0.001$), and these peak-gene pairs served as the positive examples. To generate a distance-matched set of negative examples, we determined the number of positive examples in each genomic distance bin and selected an equal number of non-significant eQTL variant-gene pairs with the highest *P* values for that bin.

Chromatin interactions evaluations

We obtained Hi-C data for cell types matching the 10x Multiome, SNARE-seq, and sci-CAR datasets[83–85]. We used 5C data for the SHARE-seq dataset[86] due to the unavailability of mouse skin Hi-C data. Chromatin interactions were mapped to a peak-gene pair by intersecting the genomic windows in the chromatin interactions data with the peak and the TSS of the gene. For Hi-C data, peak-gene pairs that were associated with the top 1% of chromatin contact frequencies were treated as positive examples, while the remaining peak-gene pairs were considered negative examples. For the 5C dataset, which evaluated the statistical significance of chromatin loops, peak-gene pairs that were associated with significant looping interactions ($P < 0.001$) were treated as positive examples, and the remaining peak-gene pairs were considered negative examples.

GrID-Net vs. correlation peak-gene prioritization

For the human corticogenesis dataset, we averaged the rankings of all candidate peak-gene 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[29].

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 [25]. 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[37,38]. 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 $\widehat{d_i} =$

594 $max(log_{10}(d_i) - 4,0)$, where d_i is the genomic distance in base pairs (bp) between the center
595 of the peak and the TSS of the gene in the peak-gene pair. This formulation assumes that only
596 distal interactions at genomic distances greater than 10 kb are affected by the genome
597 distance-dependent TF regulatory effect.
598 Positive and negative examples were chosen as described in "eQTL evaluations". We
599 held out one-third of the example as a test set, and then performed nested 3-fold cross
600 validation on the remaining examples, selecting the model that yielded the highest AUROC
601 across the folds. This model increased the AUROC on the test set compared to the predictions
602 from the averaged GrID-Net rankings (**Figure S3a**). We applied this pre-trained model on the
603 human corticogenesis dataset and found that this model outperformed the baseline predictions
604 from the averaged GrID-Net rankings (**Figure S3b**). These results demonstrated the
605 effectiveness of incorporating genomic distance as an additional feature alongside scores
606 learned from the single-cell multimodal data.
607
608 **Linking SCZ GWAS Variants to Genes**
609 We obtained FINEMAP posterior probabilities (*PP*) calculated in the original SCZ GWAS
610 study[36]. We retained all fine-mapped variants with $PP \geq 0.05$ that overlapped peaks from the
611 human corticogenesis dataset. We identified the peak that matched each SCZ variant and,
612 using the genomic distance-dependent GrID-Net scores, we linked the variant to up to 3 of the
613 peak's highest scoring genes, retaining only peak-gene pairs in the top 50% of all peak-gene
614 scores.
615
616 **Brain-specific gene expression**
617 We obtained consensus tissue-specific transcript levels from the Human Protein Atlas[87,88]. For a
618 given gene, we first calculated the mean and standard deviation of the gene's transcript levels

619 across all non-brain tissues. The brains-specific tissues were identified to be those from the
620 basal ganglia, cerebellum, cerebral cortex, medulla oblongata, midbrain, pons, and white
621 matter. We then converted the transcript level for the gene in each of the brain-specific tissues
622 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
623 level for the gene in a brain-specific tissue). Genes that had a z-score over 1 in any of the brain-
624 specific tissues were annotated as having brain-specific gene expression.
625
626 **TF binding affinity changes**
627 We obtained the reference and SCZ risk alleles for all SCZ variants from the original SCZ
628 GWAS study[36]. We then flanked each of the alleles with the genomic sequences $\pm$20 bp of the
629 SCZ variant position using the GRCh38/hg38 reference genome[89], generating a reference
630 sequence and SCZ variant sequence. We applied FIMO[90] to each of these sequences to predict
631 TF binding affinities using position weight matrices (PWMs) from HOCOMOCO v11 (Full
632 Collection)[91], JASPAR CORE (Non-Redundant) v2[92], and CIS-BP v2[93]. We evaluated the effect
633 of a specific SCZ variant on a TF's binding affinity by comparing the FIMO-calculated *p*-values
634 for the TF between the reference sequence and SCZ variant sequence.
635
636 **SCZ GPC gene expression**
637 We obtained bulk RNA-seq data from glial progenitor cells (GPCs) produced from 4 SCZ
638 patient-derived induced pluripotent cells (iPSCs) and paired controls subejcts[53,59], with at least 3
639 samples per patient. We applied a $log\ (1 + CPM)$ transformation to this data.
640 To evaluate the relationship between TF binding affinity changes caused by SCZ
641 variants to expression changes in GPCs, we first determined peak-gene pairs that are specific
642 to glial differentiation. In the human corticogenesis dataset, we identified genes that were
643 differentially expressed between glial cells and the cycling progenitor cells, which were the root

644 of the corticogenesis trajectory ($P < 0.05$, Welch's *t*-test). We then used Welch's *t*-test to
645 compare the expression levels of these glial differentiation-specific genes for each SCZ patient
646 against all of the control samples. For each gene, we identified the most statistically significant
647 difference so as to account for all genes that may be differentially expressed in any of the
648 patients compared to the controls.
649
650 **SCZ GPC co-expression pathways**
651 We first identified a set of 26 genes linked to SCZ variants causing at least a 40-fold change in
652 binding affinity for any TF. We identified 3 sets of these SCZ genes that were correlated with
653 each other (Pearson correlation coefficient ≥ 0.6) in the SCZ GPCs and paired controls
654 (SLC12A6, DLX1, DLX2; MFAP4, STAT6, NXPH4; EPN2, P2RX7, ICA1). For each of these
655 sets of SCZ genes, we then determined gene modules that may be implicated in their
656 dysregulation by identifying genes that are highly co-expressed (Pearson correlation coefficient
657 ≥ 0.8) with any of the SCZ genes in the set. We applied gene set enrichment analysis using
658 Enrichr to detect biological processes that were enriched in each of these modules.
659

660

661 **References**

662 1. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory
663 Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535–1548.e16 (2018).
664 2. Papp, B. & Plath, K. Epigenetics of Reprogramming to Induced Pluripotency. *Cell* vol. 152
665 1324–1343 Preprint at https://doi.org/10.1016/j.cell.2013.02.043 (2013).
666 3. Pierce, S. E., Granja, J. M. & Greenleaf, W. J. High-throughput single-cell chromatin
667 accessibility CRISPR screens enable unbiased identification of regulatory networks in
668 cancer. *Nat. Commun.* **12**, 2969 (2021).
669 4. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to
670 Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020).
671 5. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease.
672 *Nature Reviews Genetics* vol. 16 197–212 Preprint at https://doi.org/10.1038/nrg3891
673 (2015).
674 6. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future.
675 *Philosophical Transactions of the Royal Society B: Biological Sciences* vol. 368 20120362
676 Preprint at https://doi.org/10.1098/rstb.2012.0362 (2013).
677 7. Shukla, A. & Huangfu, D. Decoding the noncoding genome via large-scale CRISPR
678 screens. *Curr. Opin. Genet. Dev.* **52**, 70–76 (2018).
679 8. Wang, Q. *et al.* A Bayesian framework that integrates multi-omics data and gene networks
680 predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* **22**, 691–699 (2019).
681 9. Fulco, C. P. *et al.* Activity-by-contact model of enhancer-promoter regulation from
682 thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
683 10. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature*
684 vol. 593 238–243 Preprint at https://doi.org/10.1038/s41586-021-03446-x (2021).
685 11. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of

686 single cells. *Science* **361**, 1380–1385 (2018).
687 12. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and
688 chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
689 13. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. Computational principles and
690 challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
691 14. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across
692 modalities. *Nat. Methods* **17**, 14–17 (2020).
693 15. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-
694 associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).
695 16. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-
696 seq data. *Genome Biol.* **20**, 241 (2019).
697 17. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing
698 imputation methods. *Genome Biol.* **21**, 218 (2020).
699 18. Granger, C. W. J. Investigating Causal Relations by Econometric Models and Cross-
700 spectral Methods. *Econometrica* vol. 37 424 Preprint at https://doi.org/10.2307/1912791
701 (1969).
702 19. Shojaie, A. & Fox, E. B. Granger Causality: A Review and Recent Advances. *Annual*
703 *Review of Statistics and Its Application* vol. 9 289–319 Preprint at
704 https://doi.org/10.1146/annurev-statistics-040120-010930 (2022).
705 20. Singh, R., Hie, B. L., Narayan, A. & Berger, B. Schema: metric learning enables
706 interpretable synthesis of heterogeneous single-cell modalities. *Genome Biol.* **22**, 131
707 (2021).
708 21. Trevino, A. E. *et al.* Chromatin and gene-regulatory dynamics of the developing human
709 cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069.e23 (2021).
710 22. Singh, R. & Berger, B. Deciphering the species-level structure of topologically associating
711 domains. *bioRxiv* 2021.10.28.466333 (2021) doi:10.1101/2021.10.28.466333.

712 23. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-
713 associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).
714 24. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and
715 Chromatin. *Cell* vol. 183 1103–1116.e20 Preprint at
716 https://doi.org/10.1016/j.cell.2020.09.056 (2020).
717 25. Zou, Z., Ohta, T., Miura, F. & Oki, S. ChIP-Atlas 2021 update: a data-mining suite for
718 exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq
719 data. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac199.
720 26. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and
721 predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
722 27. Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**,
723 157–171 (2013).
724 28. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the
725 unknown. *Genome Biol.* **22**, 108 (2021).
726 29. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
727 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–
728 15550 (2005).
729 30. Herrmann, J. C., Beagrie, R. A. & Hughes, J. R. Making connections: enhancers in cellular
730 differentiation. *Trends Genet.* **38**, 395–408 (2022).
731 31. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association
732 studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
733 32. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in
734 regulatory DNA. *Science* **337**, 1190–1195 (2012).
735 33. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence
736 from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
737 34. Malan-Müller, S. *et al.* A systematic review of genetic variants associated with metabolic

738 syndrome in patients with schizophrenia. *Schizophr. Res.* **170**, 1–17 (2016).
739 35. Lago, S. G. & Bahn, S. The druggable schizophrenia genome: from repurposing
740 opportunities to unexplored drug targets. *NPJ Genom Med* **7**, 25 (2022).
741 36. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in
742 schizophrenia. *Nature* **604**, 502–508 (2022).
743 37. Chen, C.-H. *et al.* Determinants of transcription factor regulatory range. *Nat. Commun.* **11**,
744 2472 (2020).
745 38. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq data with
746 BETA. *Nat. Protoc.* **8**, 2502–2515 (2013).
747 39. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).
748 40. Karunakaran, K. B., Chaparala, S. & Ganapathiraju, M. K. Potentially repurposable drugs
749 for schizophrenia identified from its interactome. *Sci. Rep.* **9**, 12682 (2019).
750 41. Marballi, K. K. & Gallitano, A. L. Immediate Early Genes Anchor a Biological Pathway of
751 Proteins Required for Memory Formation, Long-Term Depression and Risk for
752 Schizophrenia. *Front. Behav. Neurosci.* **12**, 23 (2018).
753 42. Meng, X. *et al.* Neurexophilin4 is a selectively expressed α-neurexin ligand that modulates
754 specific cerebellar synapses and motor functions. *Elife* **8**, (2019).
755 43. Cardano, M. *et al.* Epsins Regulate Mouse Embryonic Stem Cell Exit from Pluripotency and
756 Neural Commitment by Controlling Notch Activation. *Stem Cells Int.* **2019**, 4084351 (2019).
757 44. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes
758 at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
759 45. Barkhuizen, W., Pain, O., Dudbridge, F. & Ronald, A. Genetic overlap between psychotic
760 experiences in the community across age and with psychiatric disorders. *Transl. Psychiatry*
761 **10**, 86 (2020).
762 46. Schröder, J. *et al.* MicroRNA-138 is a potential regulator of memory performance in
763 humans. *Front. Hum. Neurosci.* **8**, 501 (2014).

764 47. Chahrour, M. H. *et al.* Whole-exome sequencing and homozygosity analysis implicate
765 depolarization-regulated neuronal genes in autism. *PLoS Genet.* **8**, e1002635 (2012).
766 48. McLaughlin, R. L. *et al.* Genetic correlation between amyotrophic lateral sclerosis and
767 schizophrenia. *Nat. Commun.* **8**, 14774 (2017).
768 49. Kim, J. *et al.* The co-regulation mechanism of transcription factors in the human gene
769 regulatory network. *Nucleic Acids Res.* **40**, 8849–8861 (2012).
770 50. Shyn, S. I. *et al.* Novel loci for major depression identified by genome-wide association
771 study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of
772 three studies. *Mol. Psychiatry* **16**, 202–215 (2011).
773 51. Fabbri, C. & Serretti, A. Role of 108 schizophrenia-associated loci in modulating
774 psychopathological dimensions in schizophrenia and bipolar disorder. *Am. J. Med. Genet.*
775 *B Neuropsychiatr. Genet.* **174**, 757–764 (2017).
776 52. Devor, A. *et al.* Genetic evidence for role of integration of fast and slow neurotransmission
777 in schizophrenia. *Mol. Psychiatry* **22**, 792–801 (2017).
778 53. Liu, Z. *et al.* Dysregulated Glial Differentiation in Schizophrenia May Be Relieved by
779 Suppression of SMAD4- and REST-Dependent Signaling. *Cell Rep.* **27**, 3832–3843.e6
780 (2019).
781 54. Borsari, B. *et al.* Enhancers with tissue-specific activity are enriched in intronic regions.
782 *Genome Res.* **31**, 1325–1336 (2021).
783 55. Chen, R. *et al.* Cell type–specific mechanism of Setd1a heterozygosity in schizophrenia
784 pathogenesis. *Science Advances* vol. 8 Preprint at https://doi.org/10.1126/sciadv.abm1077
785 (2022).
786 56. Moore, D. L., Apara, A. & Goldberg, J. L. Krüppel-like transcription factors in the nervous
787 system: novel players in neurite outgrowth and axon regeneration. *Mol. Cell. Neurosci.* **47**,
788 233–243 (2011).
789 57. Peyrot, W. J. & Price, A. L. Identifying loci with different allele frequencies among cases of

790 eight psychiatric disorders using CC-GWAS. *Nat. Genet.* **53**, 445–454 (2021).
791 58. Pinacho, R. *et al.* Analysis of Sp transcription factors in the postmortem brain of chronic
792 schizophrenia: a pilot study of relationship to negative symptoms. *J. Psychiatr. Res.* **47**,
793 926–934 (2013).
794 59. Seifert, G., Schilling, K. & Steinhäuser, C. Astrocyte dysfunction in neurological disorders: a
795 molecular perspective. *Nat. Rev. Neurosci.* **7**, 194–206 (2006).
796 60. Tanaka, A. J. *et al.* De novo mutations in PURA are associated with hypotonia and
797 developmental delay. *Cold Spring Harb Mol Case Stud* **1**, a000356 (2015).
798 61. Daniel, D. C. & Johnson, E. M. PURA, the gene encoding Pur-alpha, member of an ancient
799 nucleic acid-binding protein family with mammalian neurological functions. *Gene* **643**, 133–
800 143 (2018).
801 62. Barešić, A., Nash, A. J., Dahoun, T., Howes, O. & Lenhard, B. Understanding the genetics
802 of neuropsychiatric disorders: the potential role of genomic regulatory blocks. *Mol.*
803 *Psychiatry* **25**, 6–18 (2020).
804 63. Bøttger, P. *et al.* Glutamate-system defects behind psychiatric manifestations in a familial
805 hemiplegic migraine type 2 disease-mutation mouse model. *Sci. Rep.* **6**, 22047 (2016).
806 64. Pieraccioli, M. *et al.* ZNF281 inhibits neuronal differentiation and is a prognostic marker for
807 neuroblastoma. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7356–7361 (2018).
808 65. Okamoto, S.-I., Sherman, K., Bai, G. & Lipton, S. A. Effect of the ubiquitous transcription
809 factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal
810 differentiation. *Brain Res. Mol. Brain Res.* **107**, 89–96 (2002).
811 66. Pinacho, R. *et al.* Increased SP4 and SP1 transcription factor expression in the postmortem
812 hippocampus of chronic schizophrenia. *J. Psychiatr. Res.* **58**, 189–196 (2014).
813 67. Wen, Y.-D. *et al.* Genetic Profiles Playing Opposite Roles of Pathogenesis in Schizophrenia
814 and Glioma. *J. Oncol.* **2020**, 3656841 (2020).
815 68. Mäki-Marttunen, V., Andreassen, O. A. & Espeseth, T. The role of norepinephrine in the

816 pathophysiology of schizophrenia. *Neurosci. Biobehav. Rev.* **118**, 298–314 (2020).
817 69. Maletic, V., Eramo, A., Gwin, K., Offord, S. J. & Duffy, R. A. The Role of Norepinephrine
818 and Its α-Adrenergic Receptors in the Pathophysiology and Treatment of Major Depressive
819 Disorder and Schizophrenia: A Systematic Review. *Front. Psychiatry* **8**, 42 (2017).
820 70. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
821 71. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell
822 multiomics experiments. *Genome Biol.* **21**, 39 (2020).
823 72. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from
824 single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).
825 73. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger
826 causality ensembles on single-cell transcriptomics. *Cell Rep.* **38**, 110333 (2022).
827 74. Andrew, G., Arora, R., Bilmes, J. & Livescu, K. Deep Canonical Correlation Analysis. in
828 *Proceedings of the 30th International Conference on Machine Learning* (eds. Dasgupta, S.
829 & McAllester, D.) vol. 28 1247–1255 (PMLR, 17--19 Jun 2013).
830 75. Hie, B. *et al.* Computational Methods for Single-Cell RNA Sequencing. *Annual Review of*
831 *Biomedical Data Science* vol. 3 339–364 Preprint at https://doi.org/10.1146/annurev-
832 biodatasci-012220-100601 (2020).
833 76. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by
834 combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
835 77. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.
836 *Cell* **174**, 1309–1324.e18 (2018).
837 78. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression
838 data analysis. *Genome Biol.* **19**, 15 (2018).
839 79. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime
840 robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
841 80. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune

842 basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
843 81. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392
844 (2012).
845 82. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and
846 splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
847 83. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing
848 human brain. *Nature* **538**, 523–527 (2016).
849 84. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of
850 chromatin interactions. *Nature* vol. 485 376–380 Preprint at
851 https://doi.org/10.1038/nature11082 (2012).
852 85. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human
853 genome. *Nature* **489**, 57–74 (2012).
854 86. Poterlowicz, K. *et al.* 5C analysis of the Epidermal Differentiation Complex locus reveals
855 distinct chromatin interaction networks between gene-rich and gene-poor TADs in skin
856 epithelial cells. *PLoS Genet.* **13**, e1006966 (2017).
857 87. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**,
858 1260419 (2015).
859 88. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome.
860 *Protein Sci.* **27**, 233–244 (2018).
861 89. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies
862 demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864
863 (2017).
864 90. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif.
865 *Bioinformatics* **27**, 1017–1018 (2011).
866 91. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor
867 binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids*

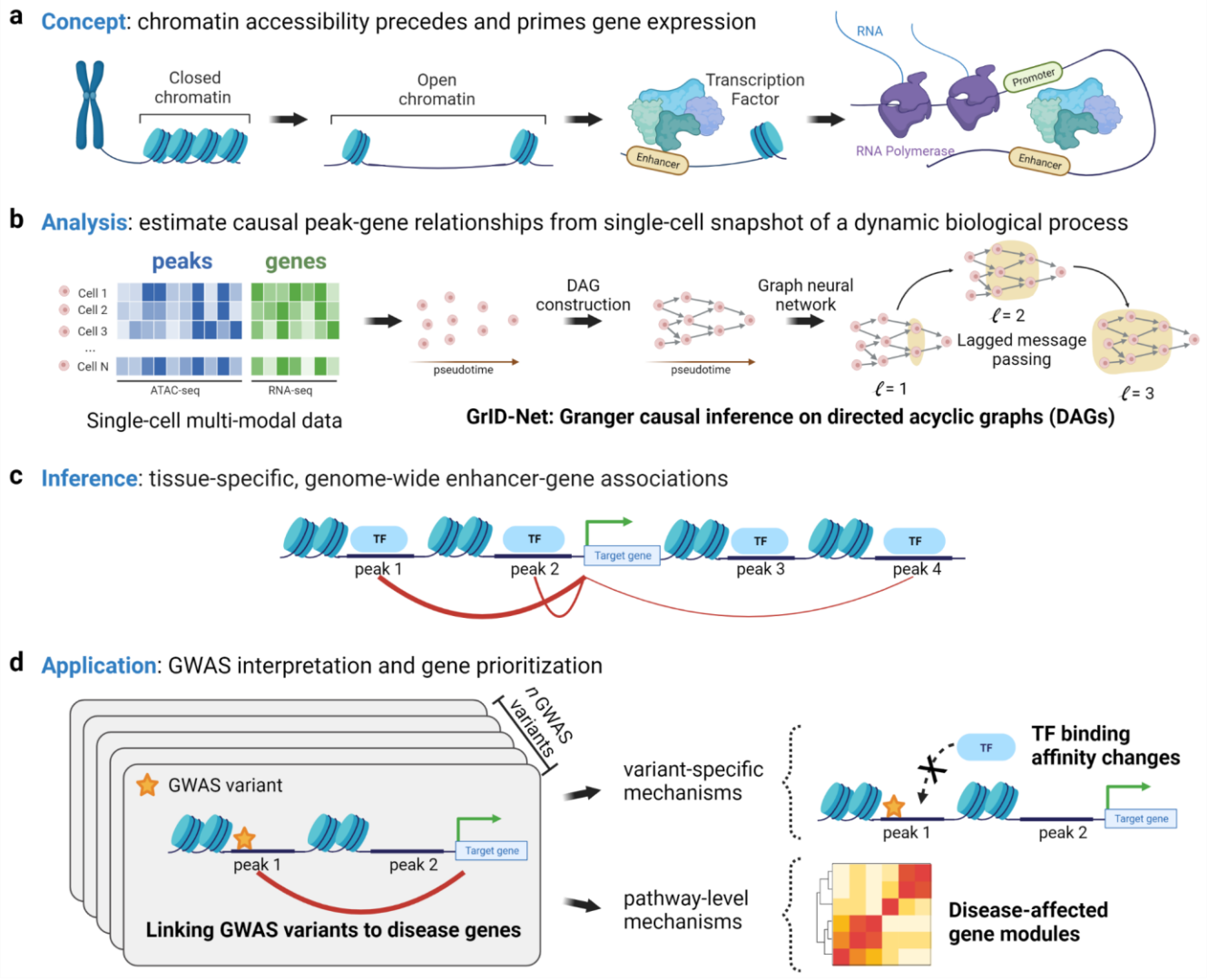
868 *Res.* **46**, D252–D259 (2018).

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

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

874

875 **Figures**



876
877

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

885 of single-cell trajectories. The core of GrID-Net is a graph neural network-based lagged
886 message passing architecture. **(c)** GrID-Net enables the detection of asynchronous interactions
887 between regulatory activity at accessible noncoding loci and their downstream effects on gene
888 expression. **(d)** GrID-Net peak-gene links enable the functional interpretation of genetic variants
889 associated with disease. We present a framework for studying the functional effects of such
890 variants both at the level of specific variants and with respect to disease-affected modules.
891

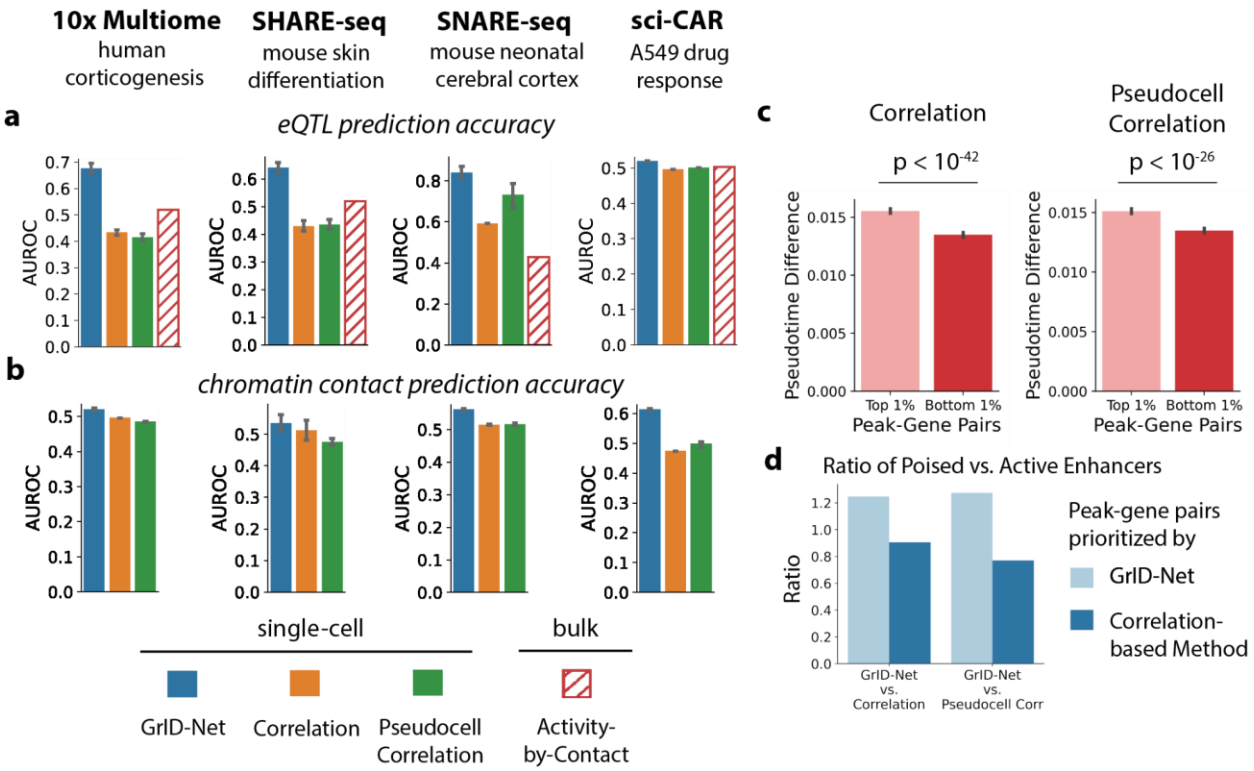


Figure 2. GrID-Net links noncoding loci to genes from single-cell multimodal dynamics. **(a)** Accuracy of single-cell (GrID-Net, Pearson correlation, and pseudocell correlation) and bulk (Activity-By-Contact) methods in predicting eQTLs. **(b)** Accuracy of methods in predicting high frequency three-dimensional 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 compared the top 1% and bottom 1% of peak-gene pairs prioritized by GrID-Net over each of the correlation-based approaches. For each peak-gene pair, we computed the separation between pseudotime values corresponding to the cells where the peak and gene achieved their maximum values, respectively. **(d)** GrID-Net effectively captures instances of chromatin priming. On the human corticogenesis dataset, we show the

906 ratio of poised enhancers to active enhancers for peaks in peak-gene pairs prioritized by GrID-

907 Net vs. each of the correlation-based methods.

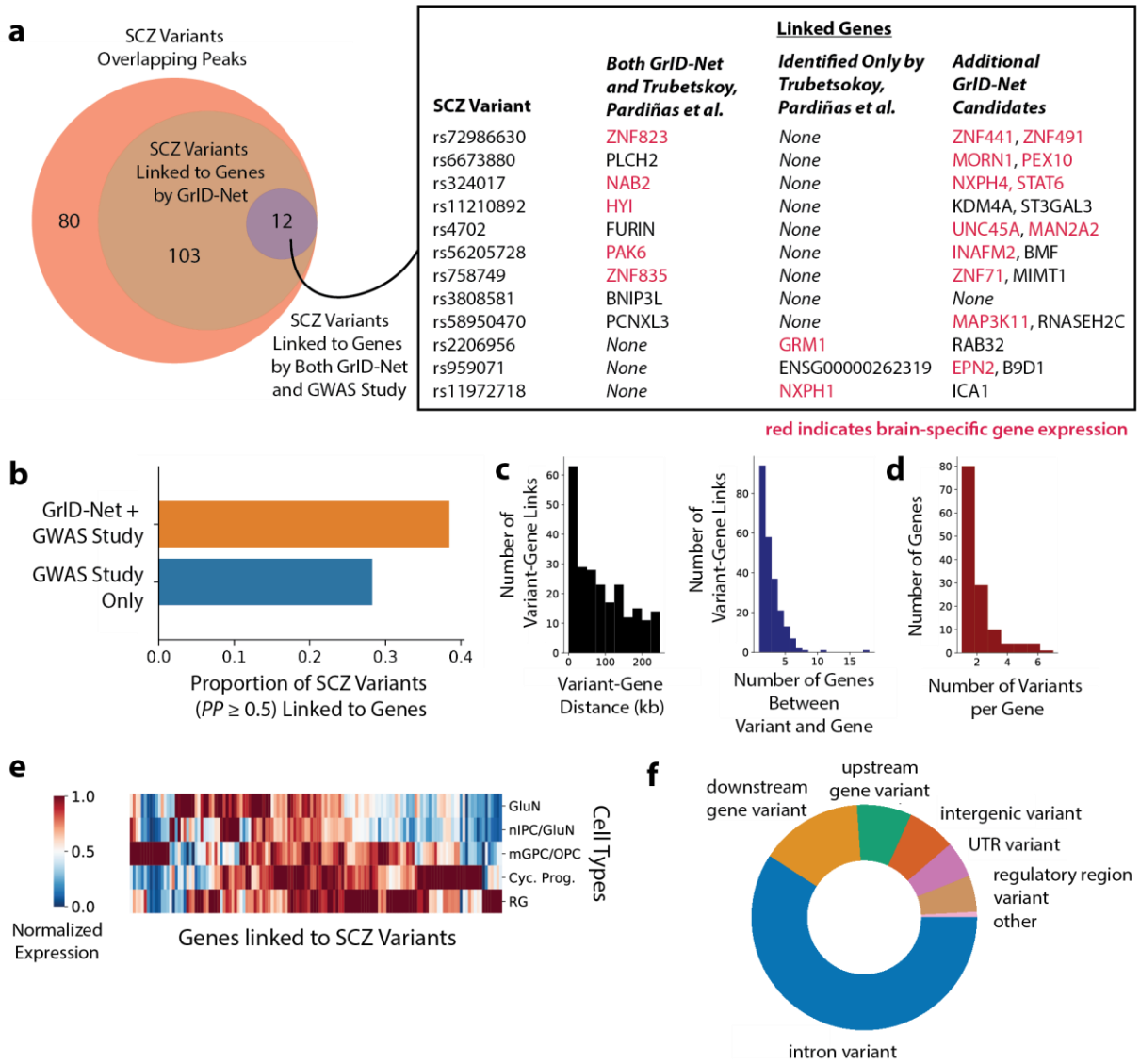


Figure 3. (a) Schizophrenia (SCZ) variants from Trubetskoy, Pardiñas *et al*., prioritized by GrID-Net. We also compared specific SCZ variant-gene links for the 12 variants that are annotated by both. **(b)** GrID-Net enables the annotation of many additional variants. As more multimodal single-cell assays become available, we expect GrID-Net's coverage to widen further. **(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

918 SCZ variants are cell type-specific. We show the heatmap of normalized expression values for
919 all genes linked to SCZ variants in cell types involved in human corticogenesis. **(f)** Distribution of
920 annotated genomic locations for SCZ variants linked to genes. A substantial fraction of genes
921 identified by GrID-Net are associated with intronic variants which have been reported to have
922 tissue-specific regulatory effects.

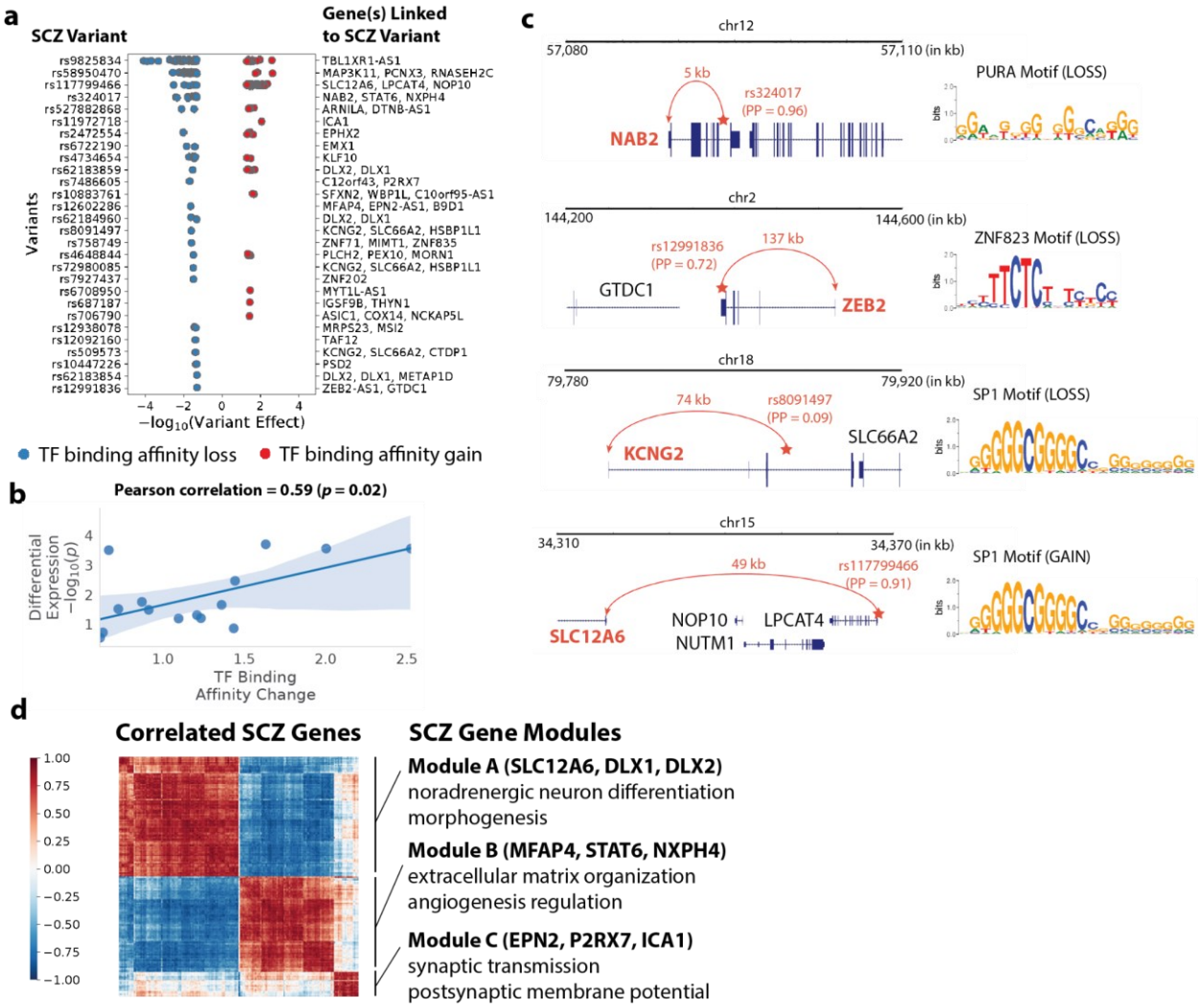


Figure 4. (a) TFs with at least a 20-fold change in predicted binding affinity caused by SCZ variants and the genes linked to these variants. Many of these TFs (e.g., KLF and SP families) are implicated in SCZ. **(b)** Relationship between TF binding affinity changes caused by SCZ variants and significance 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.

Supplementary Figures

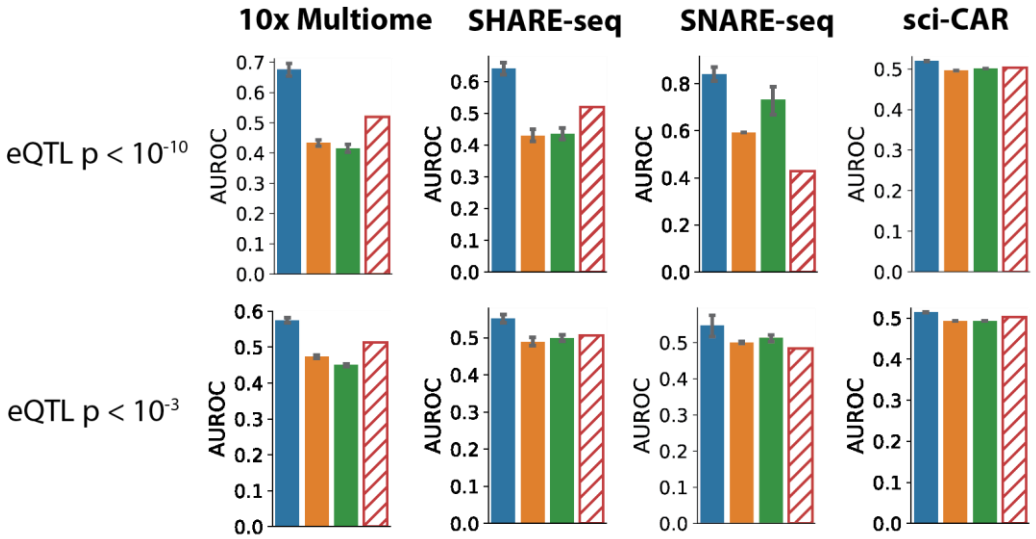


Figure S1. GrID-Net outperforms other methods in eQTL prediction accuracy across different eQTL significance thresholds.

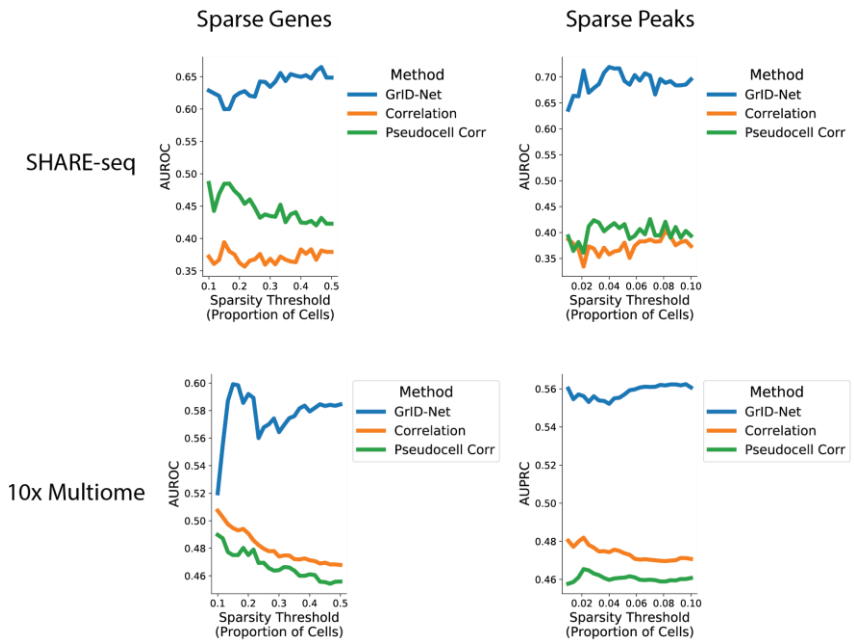


Figure S2. GrID-Net is robust to data sparsity in single-cell datasets.

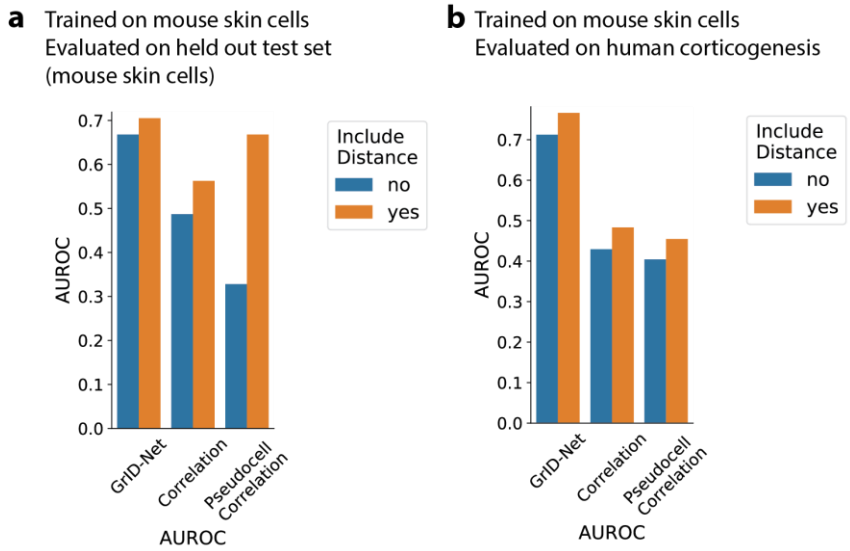


Figure S3. Genomic distance-dependent model improves peak-gene link prediction accuracy (a) within and (b) across species and cell types.

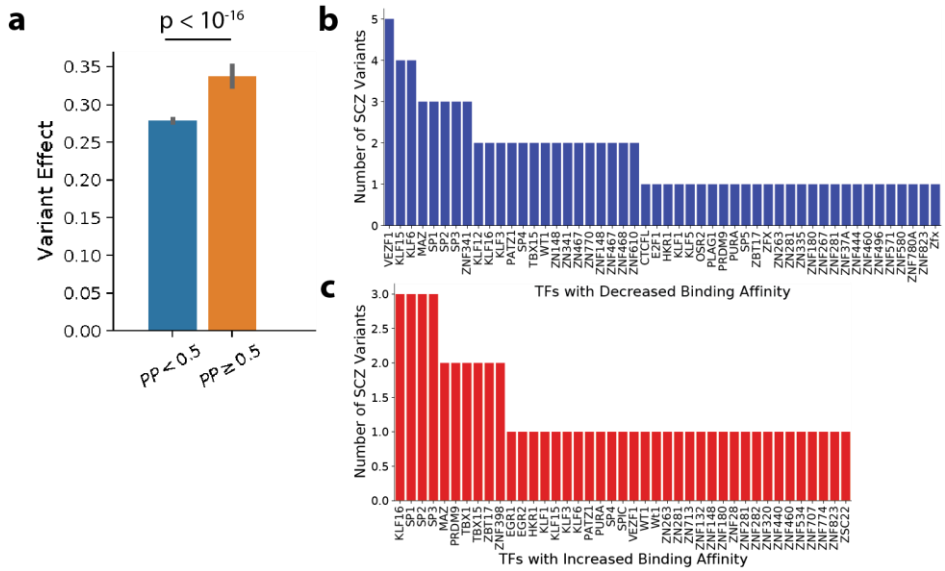


Figure S4. (a) Relationship between FINEMAP posterior probabilities (PP) for a SCZ variant and its corresponding TF binding affinity change (Welch's one-tailed *t*-test). TFs with the most widespread **(b)** increase or **(c)** decrease in binding affinity between the reference and the SCZ risk allele.