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 identify many disease-linked genetic variants at noncoding genomic loci but deciphering how these variants affect gene regulation has been a challenge. Single-cell multimodal assays that profile chromatin accessibility and gene expression in the same cell promise to reveal the causal role of noncoding loci in gene regulation. However, existing analyses have failed to bridge the gap from static single-cell snapshots to causal mechanisms. Inspired by optical parallax, which reveals object depth from static snapshots, we introduce "cell-state parallax", which reveals causal mechanisms from static single-cell snapshots by exploiting the time lag between epigenetic and transcriptional cell-states. We designed GrID-Net, a graph neural network-based generalization of Granger causal inference that leverages cell-state parallax to newly detect causal locus–gene associations in graph-structured dynamical systems, such as single-cell trajectories. We applied GrID-Net to uncover the effects of genetic variants in schizophrenia (SCZ) and uncovered noncoding mechanisms underlying the dysregulation of 132 genes, including the potassium transporters KCNG2 and SLC12A6. Furthermore, we discovered evidence for the prominent role of neural transcription-factor binding disruptions in SCZ etiology. Our work provides a general strategy for unveiling the tissue-specific impact of noncoding variants on gene dysregulation in disease and showcases

[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

the transformative potential of leveraging 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]. However, linking such noncoding loci to the genes they regulate remains a challenge. The common approach of simply associating a locus to its nearest gene is inadequate, as it neglects 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 positions of enhancers and genes and are unable to assess if a putative enhancer actually has any effect on its target gene's 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, extracting causal locus-gene associations computationally from this data remains challenging. 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 cannot capture the causal asynchrony between the chromatin and gene expression modalities of the gene regulatory

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

We refer to this asynchrony as “cell-state parallax,” drawing inspiration from the optical concept of parallax, which describes how measurements of an object from different vantage points can reveal fundamentally new insights. Cell phone cameras with dual lenses leverage optical parallax by capturing multiple 2D images of the same object from slightly offset positions to enable 3D depth perception. Analogously, in single-cell multimodal assays, we measure multiple modalities per cell, each of which captures a picture of the cell’s state but at an offset due to the time lag between the epigenome and transcriptome. By exploiting this offset, cause-effect relationships between the modalities can be uncovered. However, a fundamental challenge in accounting for these offsets is that single-cell experiments destroy cells upon measurement, so the temporal ordering of single-cell measurements is often unclear. Moreover, an inference framework that can relate measurements across cells to infer causal regulatory relationships needs to be robust to the sparsity of both scATAC-seq (< 1% non-zero entries per cell) and scRNA-seq data[16,17]. One approach to addressing sparsity is by averaging the readouts of transcriptionally-similar cells into pseudocells[11]. However, this comes at the cost of losing cell-level resolution and heterogeneity, which are crucial for discerning the subtle offsets that exist between distinct modalities in single-cell analyses.

To resolve single-cell multimodal snapshots into causal regulatory relationships, we 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 noncoding locus-gene links across a variety of cell types and single-cell multimodal assays. We then use GrID-Net to link fine-mapped genetic variants associated with schizophrenia (SCZ) to the genes they affect. We uncover noncoding mechanisms of gene dysregulation that prominently feature disruptions of neural transcription-factor binding and synaptic pathways as

putative drivers of SCZ disease pathogenesis (**Figure 1d**). Our findings demonstrate the power of GrID-Net in uncovering tissue-specific causal mechanisms of disease and development, using data generated from just one sample. Our work also illustrates the utility 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 statistical concept that can *approximate* causality in time-varying processes solely from observational data. It leans on the principle that a cause precedes its effect[18,19]. Given two time-dependent variables x (a potential cause) and y (the 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. Specifically, we generalize Granger causal inference to dynamical systems that are characterized by partial orderings. To model the asynchrony between cause and effect, we modify standard graph neural networks to allow for lagged message passing. Each node (i.e., a cell state) in the directed acyclic graph accumulates information from its parent nodes (past states), leveraging this lagged information to forecast the present state in the context 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 the candidate causal variable x. The strength of the Granger causal link between x and y can be evaluated by the statistical F-test which

102 quantifies the improvement in forecasting accuracy of the *full* model over that of the *reduced*
103 model.
104 GrID-Net enables the detection of asynchronous interactions between gene regulatory
105 elements along single-cell trajectories. For each single-cell multimodal dataset, a DAG of cell
106 states is constructed from a k-nearest neighbor graph of cells, with edges oriented in
107 accordance with pseudotime computed after a multimodal integration with Schema[20]. RNA-
108 velocity based edge orientations that capture local dynamics can also be supplied, so long as
109 any resulting cycles are broken. We evaluated the Granger causal strength between each gene
110 and chromatin accessible element ("peak") within one megabase (Mb) of the gene's
111 transcription start site, ranking the peak-gene pairs by their Granger causal strengths
112 (**Methods**).
113
114 **GrID-Net more accurately identifies noncoding locus-gene links**
115 We evaluated GrID-Net on four single-cell multimodal datasets (scATAC-seq and scRNA-seq)
116 that characterize a range of biological processes, including cell differentiation and drug
117 treatment response[11,12,21,22]. For each dataset, we sought to validate against expression
118 quantitative trait (eQTL) loci profiled in matching tissue (**Methods**). As eQTLs represent
119 associations between genetic variation at a locus and the corresponding changes in a gene's
120 expression, they serve as useful genome-wide proxies for perturbational validation of causal
121 interactions between genomic loci and genes.
122 For each gene, we intersected all ATAC-seq peaks within 1 Mb of its transcription start
123 site (TSS) with the set of eQTL variants linked to it, associating each peak with the most
124 significant variant overlapping it. We then evaluated the peak–gene pairs prioritized by GrID-Net
125 and two correlation-based methods used in previous work[11,22], Pearson correlation and
126 pseudocell correlation. The latter computes correlations on measurements aggregated across

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

152 To further benchmark GrID-Net using functional indicators of noncoding locus–gene
153 interactions, we examined three-dimensional chromatin contacts (from bulk Hi-C data) between
154 gene promoters and noncoding loci. Compared to the two correlation-based methods, we found
155 that GrID-Net more accurately prioritized peak–gene pairs that were in frequent three-
156 dimensional contact (**Figure 2b)**. The ABC method uses bulk Hi-C data itself as an input so we
157 did not evaluate its agreement with chromatin contacts. In future, the GrID-Net framework could
158 be expanded to also utilize Hi-C data when available— as we describe later, incorporating one-
159 dimensional sequence proximity as a signal into GrID-Net improves its results. However, Hi-C
160 data is currently available only for a limited set of tissues[23] and the ability to make regulatory
161 inferences from just one multimodal assay enables the convenient study of rare cell types,
162 cancer mutations and drug responses etc.
163
164 **GrID-Net reveals how chromatin accessibility primes gene activity**
165 To better understand the biological basis of GrID-Net's superior accuracy, we applied it on a
166 single-cell multimodal dataset profiling human corticogenesis. We then honed in on peak–gene
167 pairs where GrID-Net and correlation-based methods disagree. Specifically, peak–gene pairs
168 were ranked according to the difference between their GrID-Net and correlation-based
169 prioritization scores (**Methods**), with higher ranking pairs being the ones that were more
170 strongly prioritized by GrID-Net. We then investigated the distinguishing biological
171 characteristics of these differentially-prioritized pairs.
172 One of our motivations behind GrID-Net is the observation that regions of chromatin
173 around lineage-determining genes become accessible prior to gene expression, foreshadowing
174 lineage choice[21,24]. GrID-Net was designed to resolve this cell-state parallax between chromatin
175 accessibility and gene expression and we assessed if it was able to do so. We first associated
176 each peak and each gene with the pseudotime corresponding to its highest value to determine
177 the lag associated with each peak-gene pair. For both correlation-based methods, we found that

178 peak–gene pairs differentially prioritized by GrID-Net are marked by significantly longer
179 pseudotime lags between chromatin accessibility and gene expression ($P < 10^{-26}$, Welch's one-
180 tailed *t*-test, **Figure 2c**, **Methods**). This result supports GrID-Net's ability to account for the
181 asynchrony between noncoding regulatory activity and downstream gene expression control
182 that are expected to characterize causal interactions between the two. To explore the regulatory
183 implications of this asynchrony, we obtained data on histone modification states of enhancers in
184 neural progenitor cells and cross-referenced it with ATAC-seq peaks[25]. In peak–gene pairs
185 differentially prioritized by GrID-Net, more poised enhancers (H3K4me1 only) overlapped with
186 the peaks than active enhancers (both H3K4me1 and H3K27ac)[26] (**Figure 2d**, **Methods**). This
187 was not the case for peak–gene pairs differentially prioritized by correlation-based methods.
188 Poised enhancers are regulatory sites that can trigger future gene expression changes but have
189 not yet affected expression[26,27]; in contrast, active enhancers correspond to ongoing
190 transcription[28]. This suggests that resolving cell-state parallax enables the detection of lagged
191 relationships between the activities of enhancers and their target genes. GrID-Net can thus
192 leverage the concept of chromatin regulatory potential[21] to facilitate a deeper understanding of
193 tissue-specific causal regulatory mechanisms.

194 To evaluate the biological significance of capturing these putative causal interactions, we
195 examined the functions of genes implicated in peak–gene pairs differentially prioritized by GrID-
196 Net. We found these genes to be highly enriched for gene sets associated with neurogenesis
197 regulation and neural progenitor cells ($FDR < 10^{-7}$, GSEA Preranked[29]), directly relating to the
198 process of corticogenesis profiled in this dataset. We emphasize that the enrichment score is
199 calculated only on the peak–gene pairs that GrID-Net prioritizes over the correlation-based
200 methods. In comparison, the set of genes from peak–gene pairs that *all* methods score highly is
201 functionally enriched for housekeeping genes rather than neuronal activity, suggesting that
202 GrID-Net's ability to detect poised enhancers is key to discovering cell-type specific regulatory
203 relationships. Gene expression control during differentiation has been reported to depend highly

204 on asynchronous enhancer-gene interactions[30]. The enrichment of these differentiation-related
205 gene sets thus underscores GrID-Net's ability to construct accurate genome-wide maps of
206 noncoding locus–gene links that go beyond correlative analyses and towards causality.
207
208 **Linking variants to schizophrenia disease genes**
209 The ability to link noncoding loci to the genes they regulate enables the discovery of molecular
210 mechanisms underlying complex diseases. Genome-wide association studies (GWAS) have
211 identified thousands of loci associated with diseases[31]. However, over 90% of GWAS variants
212 lie in noncoding regions[32], and interpreting the functions of these noncoding variants involves
213 the challenge of linking them to the specific genes that they regulate in a cell type-specific
214 manner. With GrID-Net, we can leverage data from a single-cell assay on matching tissue to
215 uncover these cell-type-specific variant-gene links and help bridge the gap between genetic
216 variants and gene dysregulation in disease.
217 We used GrID-Net to study variants associated with schizophrenia (SCZ), a psychiatric
218 disorder for which inherited genetic variants are believed to underlie its pathogenesis[33]. SCZ's
219 high prevalence of around 1%, its substantial morbidity and mortality, and the lack of effective
220 treatments are largely due to our limited understanding of SCZ pathophysiology[34,35]. We applied
221 the GrID-Net peak-gene links inferred from human corticogenesis[22], to interpret and analyze
222 variants reported in the largest GWAS of SCZ to date[36]. We refer to this original study as the
223 Trubetskoy, Pardiñas *et al*. study.
224 To boost GrID-Net's accuracy on shorter-range peak–gene interactions, we incorporated
225 one-dimensional genomic distance as an additional feature for predicting peak–gene
226 associations. We observed a higher likelihood of peak–gene interaction if the two are located
227 close to each other, in line with reported patterns of genomic distance-dependent regulatory
228 effects[37,38]. We calibrated the relative importance of this feature on a dataset profiling mouse
229 skin differentiation[21] where we fitted a logistic regression model that combines the raw GrID-Net

230 score with a genomic distance score (**Methods**). We found that integrating genomic distance
231 with GrID-Net improved predictions of peak–gene links for mouse skin cells. The discriminative
232 power on short-range peak–gene interactions improved, while the ability to detect long-range
233 interactions was preserved. Furthermore, this model trained on mouse skin cell data performed
234 well across species and cell types in human corticogenesis as well (**Figure S3**). This genomic
235 distance-dependent model's scores were used for all remaining analyses of the human
236 corticogenesis dataset.

237 From the SCZ GWAS study, we retrieved all fine-mapped genetic variants with posterior
238 probability (*PP*) over 5%. We intersected these variants with peaks from the human
239 corticogenesis dataset. We then generated an initial set of candidate SCZ genes by associating
240 each variant with up to 3 of its overlapping peak's highest scoring genes, retaining only variant–
241 gene associations with scores in the top 50% of all peak–gene scores. This approach yielded a
242 set of 132 genes that not only is enriched for genes implicated in the original SCZ GWAS study
243 (*FDR* = 0.037, Enrichr[39]) but also includes novel SCZ gene candidates (**Supplemental Table**
244 **1**). For example, of the 12 SCZ variants that were prioritized by both GrID-Net and Trubetskoy,
245 Pardiñas *et al*., 9 were linked to the same target genes by both approaches (**Figure 3a)**. This
246 overlap included the association between rs324017 and NAB2, a gene implicated in stress-
247 induced neuroprotective maintenance that also is the downstream target of the SCZ drug
248 cinnarizine[40,41]. For these same 12 SCZ variants, GrID-Net also proposed 20 additional SCZ
249 candidate genes, of which 12 (60%) were found to be specifically expressed in brain tissues and
250 thus potentially relevant in SCZ etiology (**Methods**). Some of these additional candidates
251 include NXPH4, a neurexophilin involved in synapse function and cerebellar motor control[42],
252 and EPN2, an epsin linked to neural fate commitment[43]. Moreover, of the 78 SCZ variants with
253 *PP* over 50%, only 22 (28%) were linked to genes by the original SCZ GWAS study (**Figure 3b**).
254 Inclusion of SCZ variants that were linked to genes by GrID-Net increased this total to 30 (38%)
255 SCZ variants. GrID-Net can, therefore, be used to complement standard strategies for linking

256 variants to genes, such as those used in the original SCZ GWAS study. It can propose gene
257 associations for variants with no annotations and also augment existing annotations by
258 suggesting additional genes that a variant may be affecting.

259 In particular, GrID-Net helps address a key challenge in GWAS analyses: the common
260 strategy of assigning GWAS variants to their closest gene[44] misses many biologically
261 meaningful variant–gene relationships. By leveraging single-cell heterogeneity, GrID-Net can
262 evaluate longer range interactions. Many of the variant–gene links we identify are characterized
263 by large genomic distances between the variant and the transcription start site (TSS) of the
264 linked gene. Moreover, the variants are often linked to genes that are multiple genes over from
265 their closest gene (**Figure 3c**). For example, DCP1B and NCKAP5L, each of which are
266 connected to only a single variant, are separated from their respective variants by 18 and 11
267 genes (235 and 242 kilobases (kb)), respectively. DCP1B has been implicated in SCZ,
268 hypomania, and memory performance[45,46], while NCKAP5L, a cytoskeletal gene involved in
269 neuronal activity[47], has previously been associated with SCZ as well as amyotrophic lateral
270 sclerosis[48], indicating possible roles for these genes in neuron-specific dysregulation in SCZ.

271 In addition, 52 candidate genes (39.4% of the overall set) were associated with more
272 than one variant (**Figure 3d**). These genes may be controlled by the coordinated activity of
273 multiple regulatory elements and co-regulators relevant to SCZ[49]. IGSF9B was linked to 7
274 noncoding SCZ variants – the most of any gene. Coding variants in IGSF9B have been found to
275 be associated with major depression and the negative symptoms of SCZ[50,51], underscoring its
276 role in regulating affective symptoms in SCZ. Similarly, KCNG2, a potassium channel-encoding
277 gene, was linked to 6 noncoding SCZ variants, and has been proposed as a drug target for SCZ
278 given the known disruption of KCN-related pathways in SCZ[35,52,53]. Neither of these genes were
279 prioritized in the original GWAS analysis.

280 We also assessed the cell-type specificity of the genes linked to SCZ variants. We found
281 that 104 candidate SCZ genes (78.8% of the overall set) are differentially expressed across the

282 cell types profiled in the human corticogenesis dataset (Welch's *t*-test with Benjamini-Hochberg
283 correction $P < 0.05$, **Figure 3e**). Further analysis of the SCZ variants also revealed that 59% of
284 the variants lie in intronic regions, while only 7% of variants lie in intergenic regions more than 5
285 kb from any gene (**Figure 3f**, **Methods**). As enhancers with tissue-specific effects are highly
286 enriched in introns while intergenic enhancers more often regulate housekeeping genes[54], these
287 findings support the potential cell type-specificity of the regulatory disruptions underlying SCZ
288 pathogenesis.
289
290 **Transcription factor binding disruptions are implicated in schizophrenia**
291 Transcription factor (TF) binding at the cis-regulatory elements located in noncoding genomic
292 loci is a key mechanism of gene regulation. We observed that SCZ variants with high *PP* are
293 associated with larger TF binding affinity changes (**Figure S4a**). To investigate the
294 dysregulatory role of SCZ variants prioritized by GrID-Net, we therefore evaluated their effect on
295 TF binding. For each variant associated with a gene by GrID-Net, we calculated the difference
296 in the predicted TF binding affinity between the reference allele and the SCZ risk allele
297 (**Methods**, **Supplemental Table 2**). Of the 115 such SCZ variants, 28 (these were associated
298 with 49 target genes) were predicted to cause a 20-fold or stronger change in at least one TF's
299 binding affinity (**Figure 4a**). The TFs we identified as having the most widespread binding
300 affinity changes are known to be implicated in SCZ, exemplifying GrID-Net's ability to deliver de
301 novo mechanistic insight. VEZF1, a vasculature-related TF reported to play a role in SCZ
302 dysregulation[55], displayed diminished binding affinity most frequently (in 5 variants). In addition,
303 KLF15 and KLF6, two Kruppel-like family (KLF) TFs known to regulate neuronal growth and
304 linked to SCZ[56,36], were each associated with reduced binding affinity in 4 variants (**Figure S4b**).
305 Meanwhile, the TFs with the most widespread increase in binding affinity – SP1, SP3, and
306 KLF16 – have also been reported to be disrupted in SCZ[57,58] (**Figure S4c**). More broadly, of the
307 TFs with a 20-fold binding affinity change in at least one variant, 57.9% (=33/57) have cell type-

308 specific expression patterns in corticogenesis (Welch's *t*-test with Benjamini-Hochberg
309 correction $P < 0.05$). Collectively, these results point to genetic disruption of neuron-specific
310 regulatory elements as potential drivers of SCZ pathogenesis.

311 We reasoned that if the TF binding affinity changes at a SCZ variant are phenotypically
312 meaningful, they should result in expression changes of the corresponding gene(s) (as identified
313 by GrID-Net). Accordingly, we obtained gene expression data from glial progenitor cells (GPCs)
314 produced from SCZ patient-derived induced pluripotent cells, as well as matching controls[53,59],
315 and computed differential expression patterns in SCZ GPCs. Focusing on glial differentiation
316 genes, we observed that the differential expression of a gene in SCZ GPCs was indeed
317 significantly correlated with TF binding affinity changes at the corresponding SCZ variant
318 (Pearson correlation coefficient = 0.59, **Figure 4b**, **Methods**). The consistency between these
319 independent observations provides support to our overall analytic approach and underscores
320 the potential to leverage GrID-Net for examining causal, gene-specific regulatory mechanisms.

321 We drilled down into specific genes reported by our analysis, focusing on those linked to
322 SCZ variants associated with at least 20-fold TF binding affinity changes (**Figure 4c**). NAB2, the
323 aforementioned gene implicated in neuroprotective processes[41], was linked to rs324017, which
324 contributes to a 200-fold decline in binding affinity for the TF PURA. Intriguingly, PURA is also
325 critical for neural function and development, and its dysfunction is implicated in several
326 neurological and cognitive impairments[60,61]. In addition, rs12991836 was linked to ZEB2, whose
327 TSS is positioned 137 kb from the variant. ZEB2 has been previously reported as a target of
328 long-range regulation, and heterozygous ZEB2 mutations are known to cause Mowat-Wilson
329 syndrome, which features structural brain abnormalities and intellectual disability[62].

330 Furthermore, GrID-Net prioritized two potassium transport genes - KCNG2 and
331 SLC12A6 - a class of genes which has been associated with compromised potassium uptake
332 when dysregulated[53]. This phenotype is characteristic of not only SCZ but also other neuron-
333 related diseases, such as Huntington's disease, multiple sclerosis, and hemiplegic migraine[59,63].

334 KCNG2 was linked to 6 SCZ variants, the second most of any gene as described earlier. Using
335 TF binding affinity changes to hone in on the most relevant variants, we identified rs8091497 as
336 the variant with the largest TF binding affinity change. rs8091497 was associated with a 40-fold
337 decline in binding affinity for SP1, which itself has been linked to behavioral abnormalities in
338 SCZ[58]. As for SLC12A6, its corresponding variant (rs117799466) increases the binding affinity
339 for many TFs with repressive capabilities. The three TFs with the greatest binding affinity gains
340 at this locus – SP1, ZN281, and MAZ – all play a role in neuronal differentiation[58,64,65], and both
341 SP1 and ZN281 have also shown increased expression levels in SCZ as well[66,67]. Combined
342 with their increased binding affinity incurred by rs117799466, the upregulation of these TFs may
343 contribute to the repression of SLC12A6, an outcome that has been previously linked to the
344 potassium uptake impairments that characterize SCZ glia[53]. Together, these findings suggest a
345 possible mechanism for SLC12A6 downregulation in SCZ that implicates enhanced binding of
346 neuron-related repressors to a SLC12A6-targeting enhancer to reduce its expression levels.
347 We then broadened our analysis to gene modules, exploring if the genes identified by
348 GrID-Net are systematically involved in SCZ-related pathways. The 14 SCZ variants that cause
349 at least 40-fold changes in TF binding affinity are linked by GrID-Net to 26 genes, whose
350 differential expression patterns we examined in SCZ GPCs. We detected 3 distinct gene
351 modules that are co-expressed with several of these genes (**Figure 4d**, **Methods**). Module A
352 (SLC12A6, DLX1, and DLX2) is enriched for genes related to noradrenergic neuron
353 differentiation (*FDR* < 0.05) and morphogenesis (*FDR* < 0.02); Module B (MFAP4, STAT6,
354 NXPH4) is enriched for extracellular matrix organization (*FDR* < 10^{-14}) and angiogenesis
355 regulation (*FDR* < 0.02) genes, and Module C (EPN2, P2RX7, ICA1) is enriched for synaptic
356 transmission (*FDR* < 10^{-4}) and postsynaptic membrane potential (*FDR* < 0.05) (Enrichr[39]). Each
357 of these pathways has been previously associated with SCZ. Noradrenergic neurons have long
358 been implicated in disrupted neuromodulation in SCZ[68], neuronal angiogenesis underlies the
359 vascular abnormalities observed in SCZ[69], and genetic studies have pointed to the critical role of

synaptic function in SCZ[36]. These pathways, therefore, enable the interpretation of SCZ variants beyond the dysregulation of the specific genes that they directly affect, presenting an opportunity to discover systems-level mechanisms of SCZ pathogenesis.

Discussion

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

When studying a particular disease, GrID-Net should be paired with a single-cell assay of a matching tissue undergoing a dynamic biological process, such as development, differentiation, disease, or drug response. GrID-Net can then translate the results of this single assay to genome-wide causal hypotheses specific to the disease. We demonstrate this by leveraging a single-cell assay of human corticogenesis to interpret the functional effects of schizophrenia (SCZ) variants. GrID-Net expands both the breadth and depth of existing gene prioritization analyses: it finds novel associations for many unlinked variants while also uncovering additional causal mechanisms for previously investigated variants. For instance, we identified SCZ variants causing large (>20-fold) changes in TF binding affinity and connected them to 49

386 genes. Many of these genes are located distally from the SCZ variants that are predicted to
387 affect them and would have been neglected by existing GWAS variant–gene associations
388 analyses that consider only genomic proximity. Furthermore, by correlating TF binding affinity
389 changes with differential expression in SCZ glial progenitor cells, we predicted possible
390 mechanisms of TF-driven dysregulation for genes like KCNG2 and SLC12A6. Combined with an
391 exploration of SCZ-related pathways associated with GrID-Net-identified genes, these analyses
392 provide a comprehensive strategy for unveiling mechanistic insights about SCZ pathogenesis.
393 Such analyses can also be applied to investigate other diseases.

394 Our work rests on two conceptual insights that are broadly applicable. First, when
395 different modalities are measured in the same cell their causal relationship implies a temporal
396 offset between them, a phenomenon we refer to as “cell-state parallax”. Here, we exploit the
397 parallax between RNA-seq and ATAC-seq, but such parallax can also be leveraged from other
398 co-assayed modalities such as chromatin structure or methylation. We view RNA velocity[70] and
399 protein velocity[71] as specific applications of the broader concept of cell-state parallax, the
400 distinction being that velocity uses pre-defined mappings between modalities (e.g., spliced vs.
401 unspliced transcript-counts for a particular gene) to infer the differentiation direction of each cell.
402 In contrast, such mappings between modalities may be unclear in general cell-state parallax
403 and require customized inference approaches. Accordingly, our second conceptual insight is to
404 develop computational frameworks that can accommodate the cross-modality dynamics of cell
405 differentiation. Much of the existing work on modeling gene expression dynamics utilizes
406 mathematical and statistical tools designed for time-series analysis[72,73], which make the
407 unrealistic assumption that a total ordering of cells can capture the underlying biology. We
408 argue that a partial ordering of cells is necessary to accurately model forks and merges in
409 trajectories and accordingly formulate the differentiation landscape as a directed acyclic graph
410 of cells that can represent such an ordering. We expect a promising direction of future research

411 will be computational inference techniques, like GrID-Net, that can operate under this more
412 realistic model.
413 Since GrID-Net operates on observational data, the locus-gene associations it infers are
414 causal hypotheses that need to be validated by perturbations. In the eQTL-based evaluation, we
415 sought to mimic such validations by investigating random genetic variants observed in the
416 population and cross-referencing them to observed expression changes. We also note that such
417 causal interactions are likely indirect and mediated by additional biophysical events, such as
418 transcription factor binding. Nonetheless, indirect causality is biologically and therapeutically
419 valuable, as the underlying mechanisms can still be exploited via interventions to achieve the
420 desired phenotype.
421 With the growing availability of single-cell multimodal datasets, our approach is positioned to
422 serve as a key framework for unlocking new mechanisms of gene regulatory dynamics that
423 underlie fundamental biology and disease. Moreover, GrID-Net can be applied broadly to other
424 domains featuring graph-based dynamics, such as uncovering the drivers of social media
425 disinformation cascades. A key technical advance of this work is generalizing the applicability of
426 standard Granger causal analysis to a wider set of domains than previously possible. Doing so
427 required the use of graph-theoretic deep learning techniques to implement the core statistical
428 intuition of Granger causality for any DAG-structured ordering of observations. Our work, like
429 others before us[74], thus highlights how classical statistical concepts can be reimagined with
430 modern deep learning techniques to dramatically widen their applicability.
431
432
433

434 **Methods**

435

436 **GrID-Net formulation**

437 See Supplementary Note.

438

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

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

446

447 **Single-cell multimodal datasets**

448 We used four single-cell multimodal datasets that spanned a range of different multimodal
449 profiling techniques: 1) a 10x Multiome dataset profiling human corticogenesis, 2) a SHARE-seq
450 dataset profiling mouse skin differentiation, 3) a SNARE-seq dataset profiling mouse neonatal
451 cerebral cortical cells, 4) a sci-CAR dataset profiling A549 cells treated with dexamethasone.
452 For each of these datasets, we followed recommended practice[75] and applied $log\ (1 +$
453 $CPM/100)$ and $log\ (1 + CPM/10)$ transformations to the RNA-seq and ATAC-seq counts data,
454 respectively, where CPM indicates counts per million. We then filtered out all genes that were
455 represented in fewer than 1% of cells in the dataset and all peaks that were represented in
456 fewer than 0.1% of cells. We assembled a set of candidate peak-gene pairs by selecting all
457 peaks with centers within 1 Mb from each gene's TSS. For the 10x Multiome dataset, we only
458 considered peak-gene pairs that were separated by 250 kb to cohere with the genomic distance

threshold used in the original study[22]. 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.

485 *Pseudocell correlation*. For each sample, we identified the 49 nearest neighbors for each cell
486 based on the Euclidean distances between the joint representations learned from Schema (see
487 "Directed acyclic graph construction"). We constructed a "pseudocell" from each cell and its 49
488 nearest neighbors, constructing its RNA-seq and ATAC-seq profiles by summing the raw counts
489 from each of the modalities for all cells included in the pseudocell. We then applied $log\ (1 +$
490 $CPM/100)$ and $log\ (1 + CPM/10)$ transformations to the RNA-seq and ATAC-seq counts data,
491 respectively. Pearson correlation was applied to the transformed ATAC-seq and RNA-seq
492 values for each candidate peak-gene pair, and the absolute value of the correlation coefficients
493 were used to score the peak-gene pairs.
494
495 *Activity-by-contact (ABC) model*. We obtained ABC prediction scores in cell types matching the
496 cell types profiled in the single-cell multimodal studies. For each single-cell study, we
497 intersected the candidate peak-gene pairs with the ABC enhancer-gene pairs, assigning each
498 peak-gene pair the corresponding ABC score. If no ABC enhancer-gene pair overlap was
499 identified, the peak-gene pair was assigned a value of 0, as the ABC prediction scores were
500 available only for enhancer-gene pairs that exceeded the threshold of 0.015 defined in the
501 original study[10]. For single-cell studies that profiled mouse cells, we mapped ABC enhancer-
502 gene pairs in matching human cell types to the mouse genome due to the strong conservation
503 of epigenomes between human and mouse[80,81].
504
505 **eQTL evaluations**
506 We obtained eQTL data from the eQTL Catalogue[82] and used eQTL data specific to the cell
507 type profiled in each single-cell multimodal study. For mouse tissues, we mapped eQTL results
508 for matching human tissues to the mouse genome, as human and mouse epigenomes show
509 strong conservation[80,81].

510 We first intersected eQTL variant-gene pairs with peak-gene pairs from the single-cell
511 data and assigned each remaining peak-gene pair the most statistically significant eQTL. We
512 then binned these peak-gene pairs according to the genomic distance between each gene's
513 TSS and each peak's center, creating a bin for each of the following ranges of genomic
514 distances: [1 bp, 1 kb), [1 kb, 10kb), [10kb, 100kb), [100kb, 1 Mb). We identified all peak-gene
515 pairs that overlapped significant eQTL variant-gene pairs ($P < 0.001$), and these peak-gene
516 pairs served as the positive examples. To generate a distance-matched set of negative
517 examples, we determined the number of positive examples in each genomic distance bin and
518 selected an equal number of non-significant eQTL variant-gene pairs with the highest *P* values
519 for that bin.
520
521 **Chromatin interactions evaluations**
522 We obtained Hi-C data for cell types matching the 10x Multiome, SNARE-seq, and sci-CAR
523 datasets[83–85]. We used 5C data for the SHARE-seq dataset[86] due to the unavailability of mouse
524 skin Hi-C data. Chromatin interactions were mapped to a peak-gene pair by intersecting the
525 genomic windows in the chromatin interactions data with the peak and the TSS of the gene. For
526 Hi-C data, peak-gene pairs that were associated with the top 1% of chromatin contact
527 frequencies were treated as positive examples, while the remaining peak-gene pairs were
528 considered negative examples. For the 5C dataset, which evaluated the statistical significance
529 of chromatin loops, peak-gene pairs that were associated with significant looping interactions (P
530 < 0.001) were treated as positive examples, and the remaining peak-gene pairs were
531 considered negative examples.
532
533 **GrID-Net vs. correlation peak-gene prioritization**
534 For the human corticogenesis dataset, we averaged the rankings of all candidate peak-gene
535 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 $\widehat{d_i} =$

561 $max(log_{10}(d_i) - 4,0)$, where d_i is the genomic distance in base pairs (bp) between the center
562 of the peak and the TSS of the gene in the peak-gene pair. This formulation assumes that only
563 distal interactions at genomic distances greater than 10 kb are affected by the genome
564 distance-dependent TF regulatory effect.
565 Positive and negative examples were chosen as described in "eQTL evaluations". We
566 held out one-third of the example as a test set, and then performed nested 3-fold cross
567 validation on the remaining examples, selecting the model that yielded the highest AUROC
568 across the folds. This model increased the AUROC on the test set compared to the predictions
569 from the averaged GrID-Net rankings (**Figure S3a**). We applied this pre-trained model on the
570 human corticogenesis dataset and found that this model outperformed the baseline predictions
571 from the averaged GrID-Net rankings (**Figure S3b**). These results demonstrated the
572 effectiveness of incorporating genomic distance as an additional feature alongside scores
573 learned from the single-cell multimodal data.
574
575 **Linking SCZ GWAS Variants to Genes**
576 We obtained FINEMAP posterior probabilities (*PP*) calculated in the original SCZ GWAS
577 study[36]. We retained all fine-mapped variants with $PP \geq 0.05$ that overlapped peaks from the
578 human corticogenesis dataset. We identified the peak that matched each SCZ variant and,
579 using the genomic distance-dependent GrID-Net scores, we linked the variant to up to 3 of the
580 peak's highest scoring genes, retaining only peak-gene pairs in the top 50% of all peak-gene
581 scores.
582
583 **Brain-specific gene expression**
584 We obtained consensus tissue-specific transcript levels from the Human Protein Atlas[87,88]. For a
585 given gene, we first calculated the mean and standard deviation of the gene's transcript levels

586 across all non-brain tissues. The brains-specific tissues were identified to be those from the
587 basal ganglia, cerebellum, cerebral cortex, medulla oblongata, midbrain, pons, and white
588 matter. We then converted the transcript level for the gene in each of the brain-specific tissues
589 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
590 level for the gene in a brain-specific tissue). Genes that had a z-score over 1 in any of the brain-
591 specific tissues were annotated as having brain-specific gene expression.
592
593 **TF binding affinity changes**
594 We obtained the reference and SCZ risk alleles for all SCZ variants from the original SCZ
595 GWAS study[36]. We then flanked each of the alleles with the genomic sequences $\pm$20 bp of the
596 SCZ variant position using the GRCh38/hg38 reference genome[89], generating a reference
597 sequence and SCZ variant sequence. We applied FIMO[90] to each of these sequences to predict
598 TF binding affinities using position weight matrices (PWMs) from HOCOMOCO v11 (Full
599 Collection)[91], JASPAR CORE (Non-Redundant) v2[92], and CIS-BP v2[93]. We evaluated the effect
600 of a specific SCZ variant on a TF's binding affinity by comparing the FIMO-calculated *p*-values
601 for the TF between the reference sequence and SCZ variant sequence.
602
603 **SCZ GPC gene expression**
604 We obtained bulk RNA-seq data from glial progenitor cells (GPCs) produced from 4 SCZ
605 patient-derived induced pluripotent cells (iPSCs) and paired controls subejcts[53,59], with at least 3
606 samples per patient. We applied a $log\ (1 + CPM)$ transformation to this data.
607 To evaluate the relationship between TF binding affinity changes caused by SCZ
608 variants to expression changes in GPCs, we first determined peak-gene pairs that are specific
609 to glial differentiation. In the human corticogenesis dataset, we identified genes that were
610 differentially expressed between glial cells and the cycling progenitor cells, which were the root

611 of the corticogenesis trajectory ($P < 0.05$, Welch's *t*-test). We then used Welch's *t*-test to
612 compare the expression levels of these glial differentiation-specific genes for each SCZ patient
613 against all of the control samples. For each gene, we identified the most statistically significant
614 difference so as to account for all genes that may be differentially expressed in any of the
615 patients compared to the controls.
616
617 **SCZ GPC co-expression pathways**
618 We first identified a set of 26 genes linked to SCZ variants causing at least a 40-fold change in
619 binding affinity for any TF. We identified 3 sets of these SCZ genes that were correlated with
620 each other (Pearson correlation coefficient ≥ 0.6) in the SCZ GPCs and paired controls
621 (SLC12A6, DLX1, DLX2; MFAP4, STAT6, NXPH4; EPN2, P2RX7, ICA1). For each of these
622 sets of SCZ genes, we then determined gene modules that may be implicated in their
623 dysregulation by identifying genes that are highly co-expressed (Pearson correlation coefficient
624 ≥ 0.8) with any of the SCZ genes in the set. We applied gene set enrichment analysis using
625 Enrichr to detect biological processes that were enriched in each of these modules.

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. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).
16. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).
17. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).
18. 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).
19. 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).
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. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103–1116.e20 (2020).
22. 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).

23. 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.
24. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).
25. 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.
26. 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).
27. Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
28. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol.* **22**, 108 (2021).
29. 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).
30. Herrmann, J. C., Beagrie, R. A. & Hughes, J. R. Making connections: enhancers in cellular differentiation. *Trends Genet.* **38**, 395–408 (2022).
31. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
32. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
33. 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).
34. 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).

35. Lago, S. G. & Bahn, S. The druggable schizophrenia genome: from repurposing opportunities to unexplored drug targets. *NPJ Genom Med* **7**, 25 (2022).

36. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

37. Chen, C.-H. *et al.* Determinants of transcription factor regulatory range. *Nat. Commun.* **11**, 2472 (2020).

38. Wang, S. *et al.* Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* **8**, 2502–2515 (2013).

39. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, e90 (2021).

40. Karunakaran, K. B., Chaparala, S. & Ganapathiraju, M. K. Potentially repurposable drugs for schizophrenia identified from its interactome. *Sci. Rep.* **9**, 12682 (2019).

41. 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).

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

43. 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).

44. 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).

45. 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).

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

47. Chahrour, M. H. *et al.* Whole-exome sequencing and homozygosity analysis implicate

depolarization-regulated neuronal genes in autism. *PLoS Genet.* **8**, e1002635 (2012).

48. McLaughlin, R. L. *et al.* Genetic correlation between amyotrophic lateral sclerosis and schizophrenia. *Nat. Commun.* **8**, 14774 (2017).
49. Kim, J. *et al.* The co-regulation mechanism of transcription factors in the human gene regulatory network. *Nucleic Acids Res.* **40**, 8849–8861 (2012).
50. 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).
51. 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).
52. Devor, A. *et al.* Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia. *Mol. Psychiatry* **22**, 792–801 (2017).
53. 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).
54. Borsari, B. *et al.* Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res.* **31**, 1325–1336 (2021).
55. 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).
56. 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).
57. 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).

58. 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).

59. Seifert, G., Schilling, K. & Steinhäuser, C. Astrocyte dysfunction in neurological disorders: a molecular perspective. *Nat. Rev. Neurosci.* **7**, 194–206 (2006).

60. 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).

61. 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).

62. 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).

63. 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).

64. 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).

65. 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).

66. 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).

67. Wen, Y.-D. *et al.* Genetic Profiles Playing Opposite Roles of Pathogenesis in Schizophrenia and Glioma. *J. Oncol.* **2020**, 3656841 (2020).

68. Mäki-Marttunen, V., Andreassen, O. A. & Espeseth, T. The role of norepinephrine in the pathophysiology of schizophrenia. *Neurosci. Biobehav. Rev.* **118**, 298–314 (2020).

69. 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).
70. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
71. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020).
72. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).
73. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Rep.* **38**, 110333 (2022).
74. 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).
75. 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).
76. Cusanovich, D. A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
77. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309–1324.e18 (2018).
78. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
79. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
80. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).

81. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392 (2012).

82. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).

83. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

84. 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).

85. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

86. 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).

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

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

89. 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).

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

91. 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).

92. 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).

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

Figures

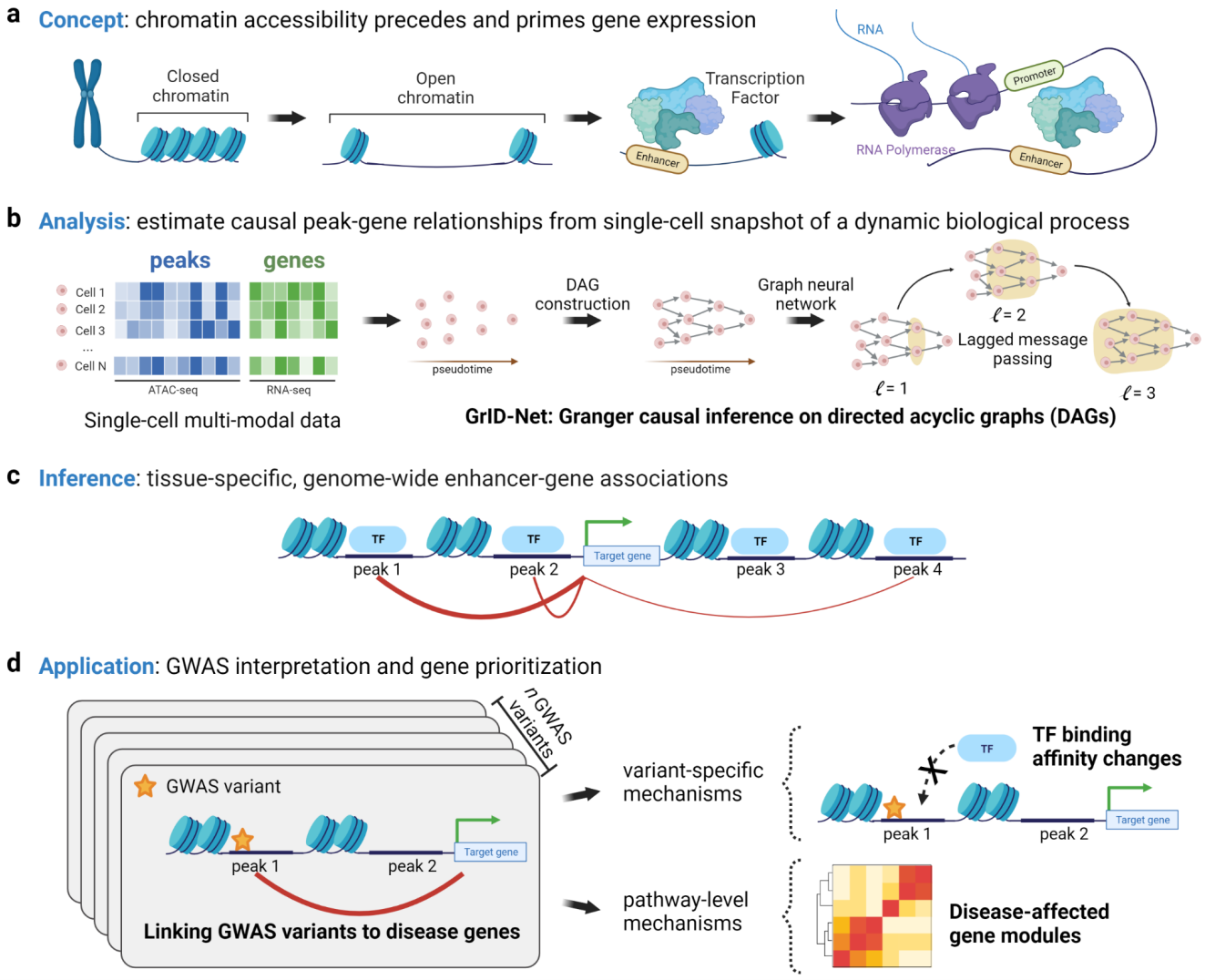

Figure 1. GrID-Net Framework. (a) The causal asynchrony between chromatin accessibility and gene expression results in "cell-state parallax", where a temporal lag exists between these two modalities. Leveraging this parallax from single-cell snapshots enables inference of causal locus-gene associations.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

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

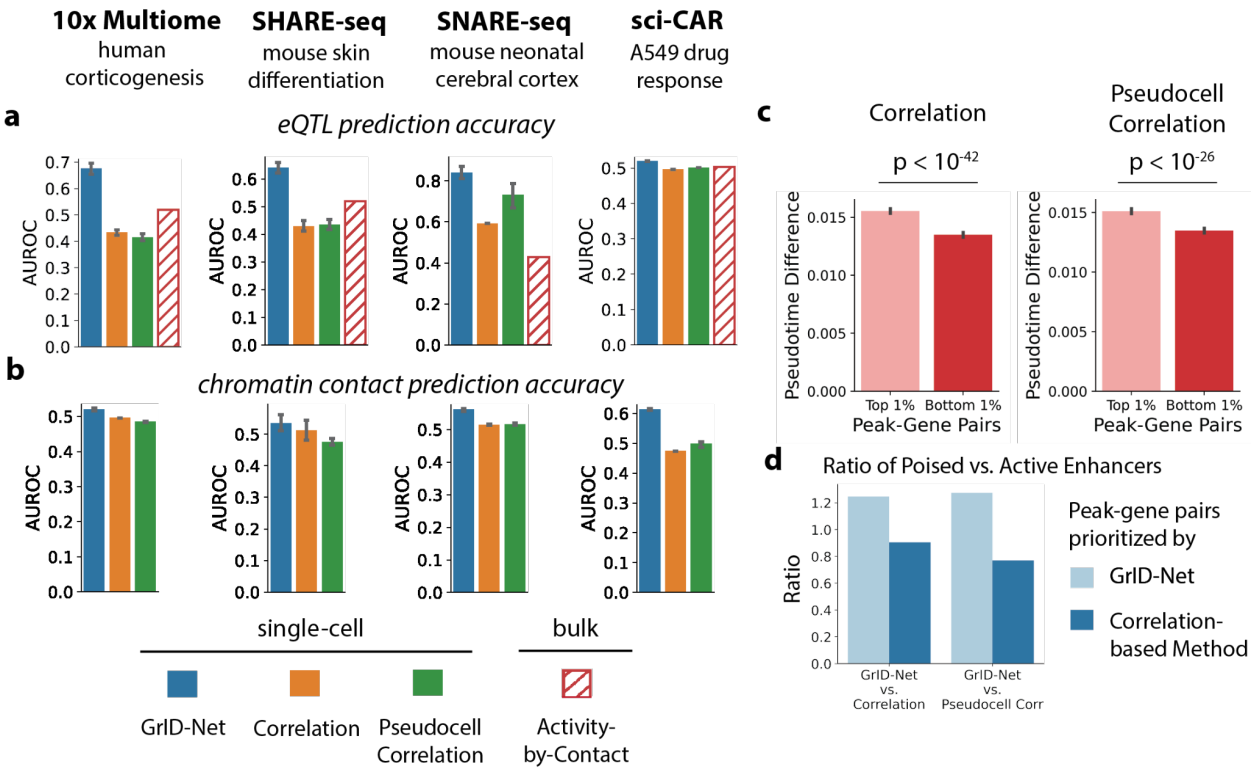

Figure 2. GrID-Net links noncoding loci to genes from single-cell multimodal dynamics. **(a)** Accuracy of single-cell (GrID-Net, Pearson correlation, and pseudocell correlation) and bulk (Activity-By-Contact) methods in predicting expression quantitative trait loci (eQTLs) in matching tissue. In this 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,

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

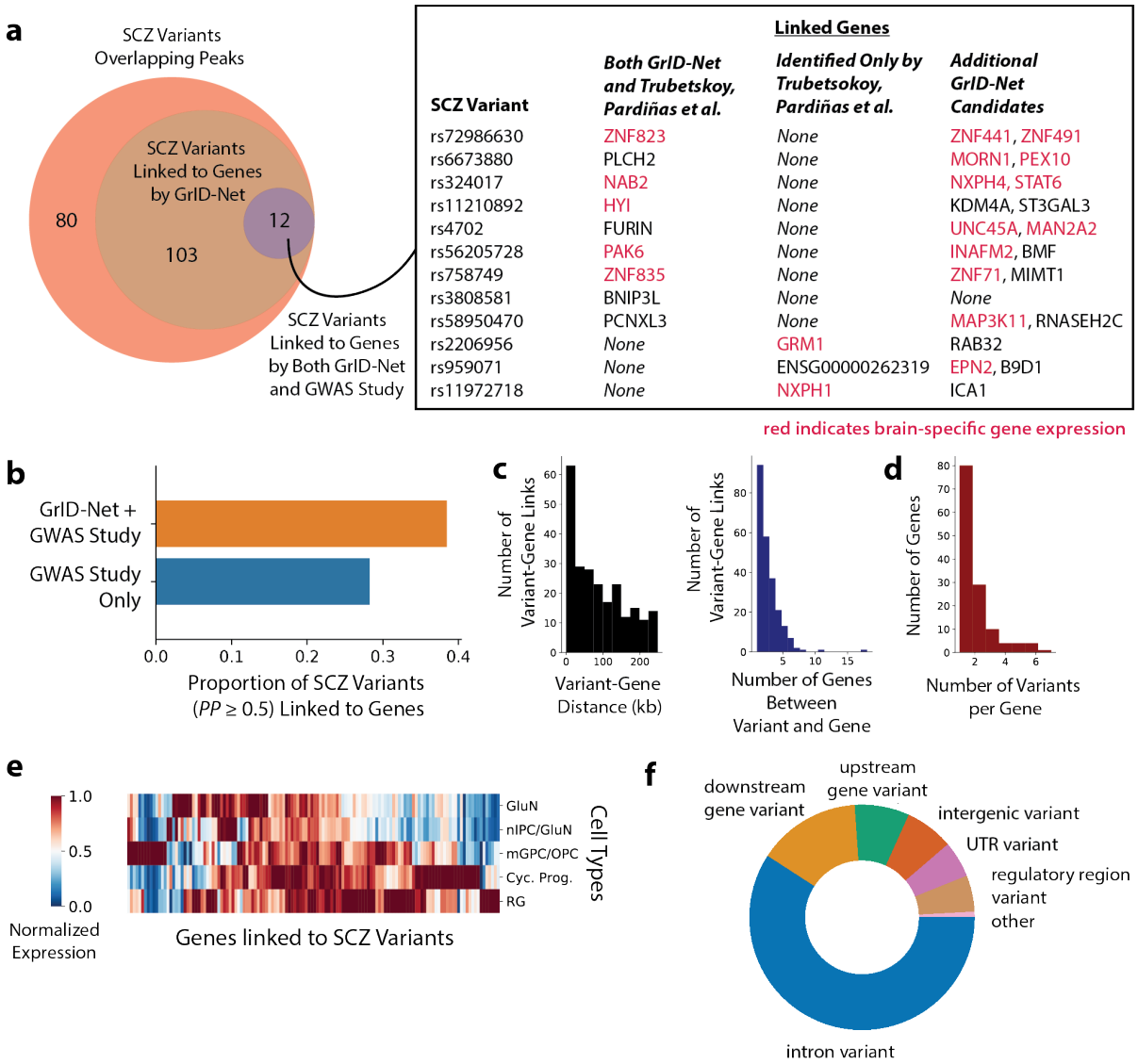

Figure 3. GrID-Net predictions link schizophrenia variants to affected genes. (a,b) Schizophrenia (SCZ) variants from Trubetskoy, Pardiñas *et al.*, prioritized by GrID-Net. Overall, 195 SCZ variants overlapped with scATAC-seq peaks from the human corticogenesis dataset. Compared to the original study, 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.

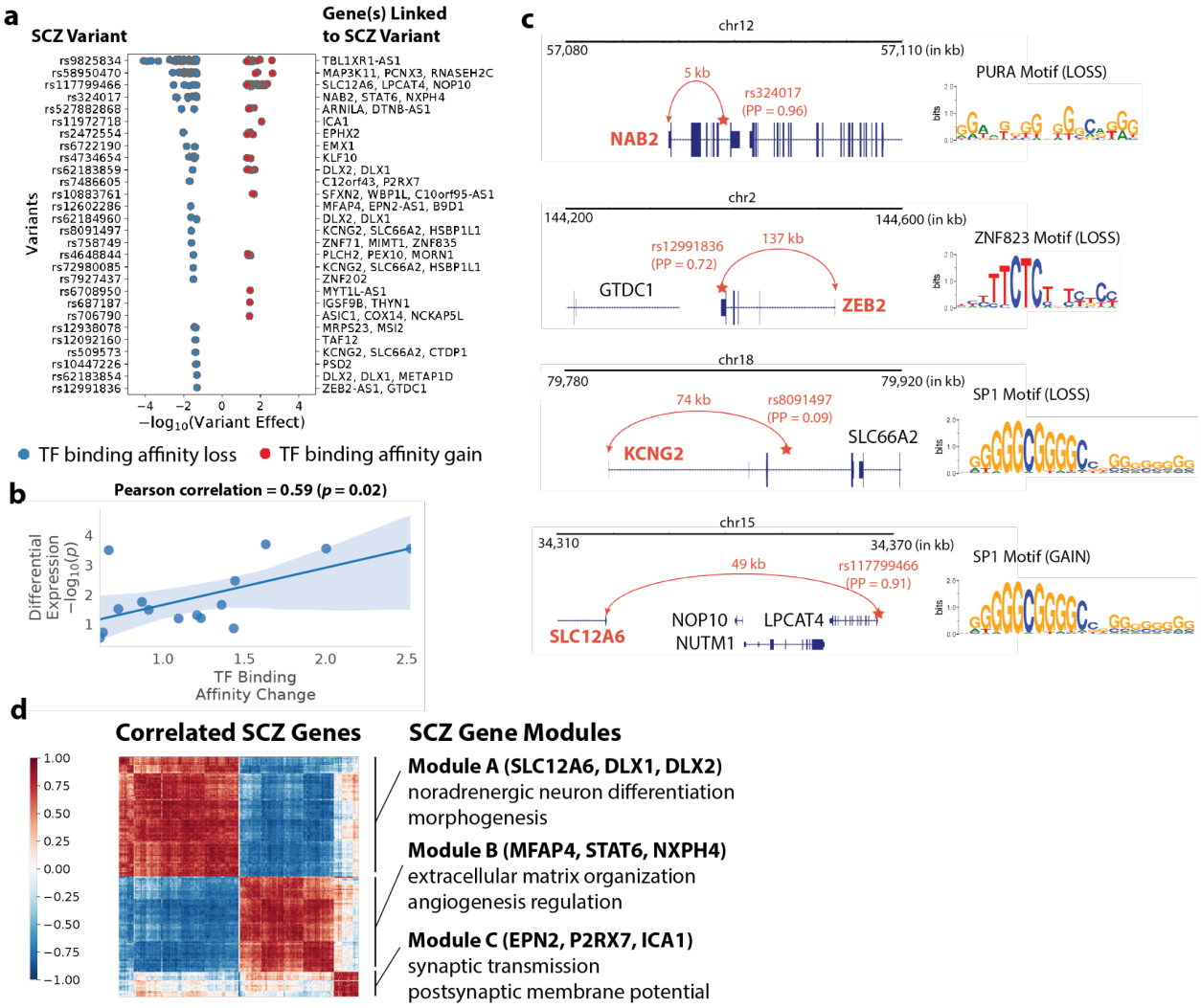

Figure 4. GrID-Net links transcription factor binding disruptions to gene dysregulation in schizophrenia. (a) TFs with at least a 20-fold change in predicted binding affinity caused by SCZ variants and the genes linked to these variants. Many of these TFs (e.g., KLF and SP families) are implicated in 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.

690

Supplementary Figures

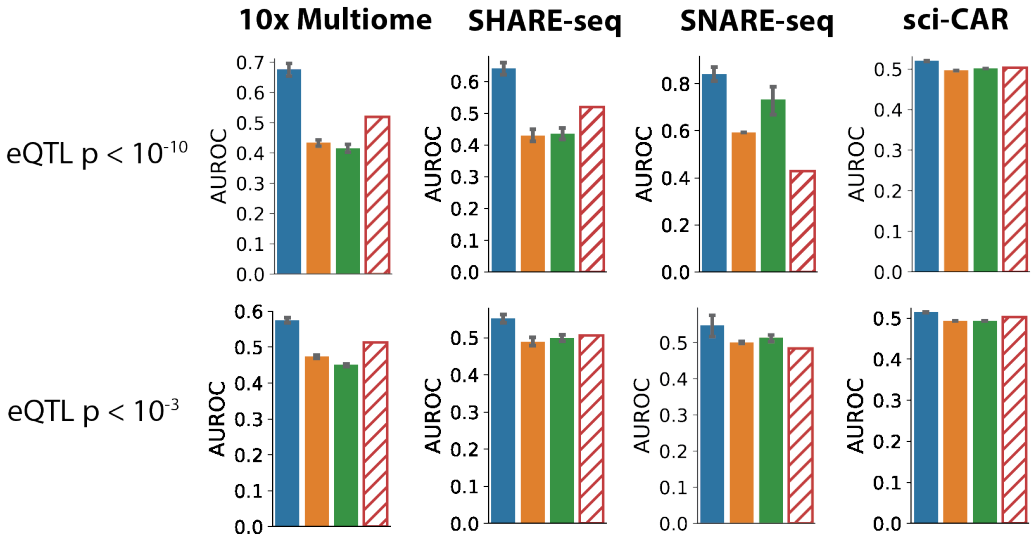

Figure S1. GrID-Net outperforms other methods in eQTL prediction accuracy across different eQTL significance thresholds.

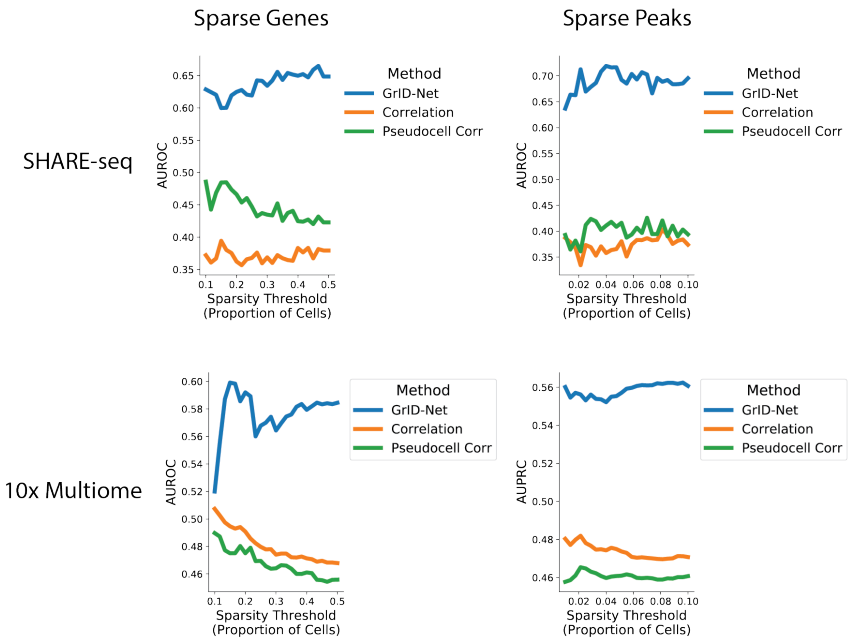

Figure S2. GrID-Net is robust to data sparsity in single-cell datasets.

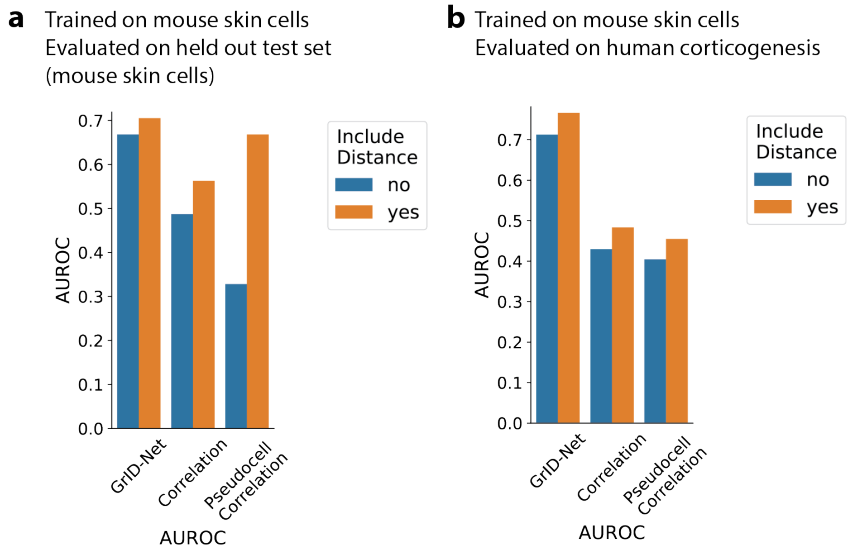

Figure S3. Genomic distance-dependent model improves peak-gene link prediction accuracy (a) within and (b) across species and cell types.

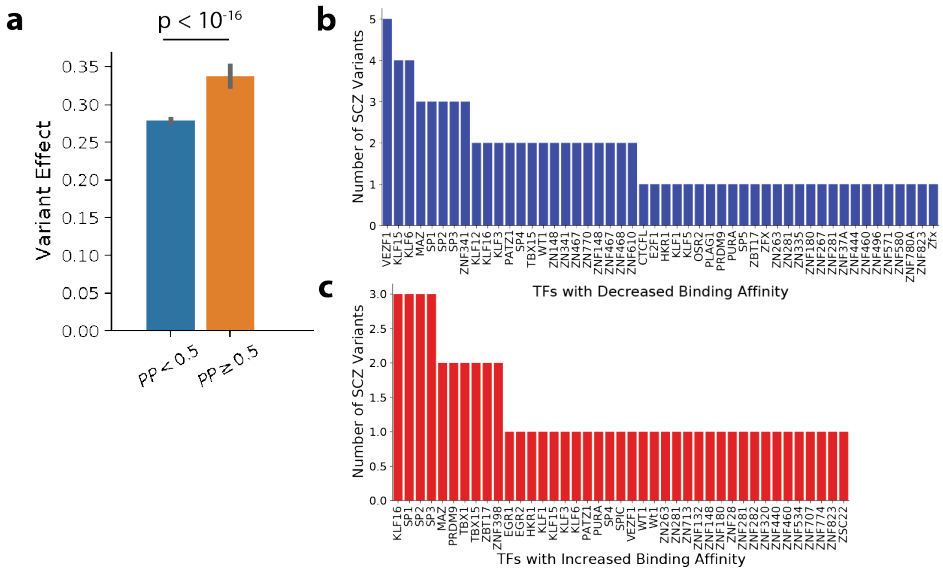

Figure S4. (a) Relationship between FINEMAP posterior probabilities (PP) for a SCZ variant and its corresponding TF binding affinity change (Welch's one-tailed *t*-test). TFs with the most widespread **(b)** increase or **(c)** decrease in binding affinity between the reference and the SCZ risk allele.

Supplementary Note

1 Background on Granger Causality: Standard Formulations

In time series analyses, the classical Granger causality framework uses the vector autoregressive (VAR) model, which can be expressed as follows (Lütkepohl, 2005).

$$y_t = \sum_{\ell=1}^{L} a_1^{(\ell)} y_{t-\ell} + \sum_{\ell=1}^{L} a_2^{(\ell)} x_{t-\ell} + \epsilon_t \tag{1}$$

Here, $y_t, x_t \in \mathbb{R}$ are values at time t for two stationary time series y and x, respectively; $a_1^{(\ell)}$ and $a_2^{(\ell)}$ are coefficients that specify how lag ℓ affects the future of time series y; and ϵ_t is a zero-mean noise term. In this model, y_t is assumed to be a linear combination of the L most recent values of y and x each, with x said to Granger cause time series y if and only if $a_2^{(\ell)} \neq 0$ for all ℓ.

A more generalized formulation — the one we follow — for performing Granger causal inference is to consider two related models for forecasting time series y. A *full* model considers the past values of both y and x to forecast y, while a *reduced* model excludes the effect of x, only containing terms related to the past values of y.

$$y_t = f^{(full)}(y_{t-1}, ..., y_{t-L}; x_{t-1}, ..., x_{t-L}) + \epsilon_t \tag{2}$$

$$y_t = f^{(reduced)}(y_{t-1}, ..., y_{t-L}) + \epsilon_t \tag{3}$$

Here, $f^{(full)}(\cdot)$ and $f^{(reduced)}(\cdot)$ are generalized functions that specify how the value of y_t depends on past values of y and x. The predictions of the two models are then compared, upon which a Granger causal relationship is declared if the full model's predictions are significantly more accurate.

2 Granger Causality on a DAG: Graph Neural Network Formulation

Let the data be represented by a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $n = |\mathcal{V}|$ nodes (i.e., observations) and directed edges $\mathcal{E}$ indicating the information flow or partial order between these observations. For instance, when applied to the case of standard Granger causal inference on time series data, $\mathcal{G}$ would be the linear graph corresponding to the time series. Let $\boldsymbol{y}, \boldsymbol{x} \in \mathbb{R}^n$ correspond to the values of variables y and x on the nodes in $\mathcal{V}$, with x putatively Granger-causing y. To infer Granger causality, we compare the full and reduced models as above:

$$\boldsymbol{y} = g(\tilde{\boldsymbol{h}}_y^{(full)} + c\tilde{\boldsymbol{h}}_x^{(full)}) + \boldsymbol{\epsilon} \tag{4}$$

$$\boldsymbol{y} = g(\tilde{\boldsymbol{h}}_y^{(reduced)}) + \boldsymbol{\epsilon} \tag{5}$$

Here, $\tilde{\boldsymbol{h}}_y^{(full)}, \tilde{\boldsymbol{h}}_y^{(reduced)}, \tilde{\boldsymbol{h}}_x^{(full)} \in \mathbb{R}^n$ represent the historical information of y or x (as denoted by the subscript) that is aggregated by graph neural network (GNN) layers. For a history $\tilde{\boldsymbol{h}}$, the value $\tilde{\boldsymbol{h}}[v]$ at node $v \in \mathcal{V}$ is the information accumulated from v's ancestors in $\mathcal{G}$. While $\tilde{\boldsymbol{h}}_y^{(full)}$ and $\tilde{\boldsymbol{h}}_y^{(reduced)}$ are outputs determined from similar architectures, their learned weights and bias terms end up being different since $\tilde{\boldsymbol{h}}_y^{(full)}$ is influenced by its interaction with $\tilde{\boldsymbol{h}}_x$. The coefficient c mediates this interaction and describes the effect of x's history on y. In addition, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a zero-mean noise term. We also include $g(\cdot)$ as an optional link function that maps its input to the support

of the distribution of variable y. In our single-cell application, we set g as the exponential function since the target variable ($\boldsymbol{y}$) represents normalized transcript counts, which are non-negative.

For brevity, we only describe how $\tilde{\boldsymbol{h}}_x^{(full)}$ is computed, with $\tilde{\boldsymbol{h}}_y^{(full)}$ and $\tilde{\boldsymbol{h}}_y^{(reduced)}$ being independently computed analogously. Also, since $\boldsymbol{x}$ occurs only in the full model, for notational convenience we drop its superscript *(full)* below, writing just $\tilde{\boldsymbol{h}}_x$. We express $\tilde{\boldsymbol{h}}_x$ as the mean of the outputs of L consecutive GNN layers, denoting the layerwise outputs as $\boldsymbol{h}_x^{(\ell)}$:

$$\tilde{\boldsymbol{h}}_x = \frac{1}{L}\sum_{\ell=1}^{L} \boldsymbol{h}_x^{(\ell)} \tag{6}$$

$$\boldsymbol{h}_x^{(\ell)} = \begin{cases} \sigma(w_x^{(\ell)} \boldsymbol{A}_+^T \boldsymbol{h}_x^{(\ell-1)} + b_x^{(\ell)}) & \text{if } \ell > 1 \\ \sigma(w_x^{(\ell)} \boldsymbol{A}^T \boldsymbol{x} + b_x^{(\ell)}) & \text{if } \ell = 1 \end{cases} \tag{7}$$

Here, $w_x^{(\ell)}, b_x^{(\ell)} \in \mathbb{R}$ are the per-layer weight and bias terms, respectively. $\sigma(\cdot)$ represents the nonlinear activation in each of the GNN layers, chosen here to be the hyperbolic tangent function. $\boldsymbol{A}$ and $\boldsymbol{A}_+ \in \mathbb{R}^{n\times n}$ are matrices defined by the DAG structure. $\boldsymbol{A}$ is the column-normalized adjacency matrix of $\mathcal{G}$, with $A_{ij} = \frac{1}{d_j}$ if edge $(i,j) \in \mathcal{E}$ and 0 otherwise, where d_j is the in-degree of node j. We note that a DAG does not have self-loops, so the diagonal terms of $\boldsymbol{A}$ are zero. In $\boldsymbol{A}_+$, we include these diagonal terms, setting them to the same value as others in the column (i.e., d_j is incremented by 1 in $\boldsymbol{A}_+$ to preserve normalization).

Together, $\boldsymbol{A}$ and $\boldsymbol{A}_+$ allow us to extend the key intuition of Granger causality to DAGs: predicting a value at node v from the values of its ancestors that are within L steps. The first layer introduces the lag, using information at the parents of v but not the information at v itself. Each subsequent layer introduces the preceding set of parents, with the diagonal term in $\boldsymbol{A}_+$ ensuring that information already stored at v is integrated with the new information propagated via v's parents. The sequence of GNN layers therefore reflects the successive aggregation of past information by traversing L steps backward along the DAG for each node in the graph. Thus, we are able to conceptually match time series-based formulations (e.g., Eqn. 1) but with the crucial advantage of leveraging the richness of a DAG's structure in directing the information flow and aggregating sparse data.

3 Comparing Full and Reduced Models to Infer Granger Causality

Let $\Theta_y^{(full)}, \Theta_y^{(reduced)}, \Theta_x^{(full)} \in \mathbb{R}^{2L}$ denote the set of parameters for the full and reduced models, with $\Theta_x^{(full)} = \{w_x^{(1)}, ..., w_x^{(L)}; b_x^{(1)}, ..., b_x^{(L)}\}$ and $\Theta_y^{(full)}, \Theta_y^{(reduced)}$ defined analogously. All the Θ parameters are jointly learned by minimizing the combined loss of the full and reduced models (Montalto et al., 2015):

$$\mathcal{L}^{(total)} = \sum_{x,y\in\mathcal{P}} \left(\sum_{v\in\mathcal{V}} \mathcal{L}(\hat{y}_v^{(full)}, y_v) + \sum_{v\in\mathcal{V}} \mathcal{L}(\hat{y}_v^{(reduced)}, y_v) \right) \tag{8}$$

where $\hat{y}_v^{(full)}$ and $\hat{y}_v^{(reduced)}$ correspond to the predictions for observation v by the full and reduced models, respectively. We note that the full and reduced models have completely separate parameters. Also, in a multivariate system, not all pairwise combinations of variables may be relevant: $\mathcal{P}$ is the subset of x, y variable pairs whose putative Granger-causal interactions are of interest.

To infer Granger causal interactions between y and x, we compare the set of loss terms associated with the full model $\mathcal{L}(\hat{y}_v^{(full)}, y_v)$ to those of the reduced model $\mathcal{L}(\hat{y}_v^{(reduced)}, y_v)$, noting that the precise functional form of $\mathcal{L}$ depends on the domain. If x does not Granger cause y, the full and reduced models will have similar explanatory power and hence similar loss values. Otherwise, the full model will have a lower loss. We assess this by a one-tailed F-test, comparing the residual sum of squares of the reduced and full models. We rank the set of candidate x–y interactions by the F-statistic, with a higher score corresponding to stronger evidence for a Granger causal interaction.

4 DOMAIN-SPECIFIC MODEL CUSTOMIZATION

The loss $\mathcal{L}$ should be chosen as appropriate for the domain. In our single-cell context, y corresponds to gene expression. We preprocessed RNA-seq transcript counts, normalizing and log transforming them so that the mean squared error loss was appropriate; we note that most methods based on Pearson correlation of gene expression also seek to minimize squared loss, explicitly or implicitly.

The GrID-Net model can also be customized to account for different lags and lookbacks. The number of GNN layers L corresponds to the maximum amount of past information desired. Similarly, to introduce a k-hop lag, the first k GNN layers would use the matrix $\boldsymbol{A}$ while the later layers would use $\boldsymbol{A}_+$. In the sections above, we described a one-hop lag that we have used for all analyses here.

REFERENCES

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. MIT Press, 2005.

Alessandro Montalto, Sebastiano Stramaglia, Luca Faes, Giovanni Tessitore, Roberto Preveteg, and Daniele Marinazzoa. Neural networks with non-uniform embedding and explicit validation phase to assess granger causality. *Neural Networks*, 71:159–171, 2015.

	A	B	C	D	E	F	G	H	I
1	Table S1								
2	Notes: 10x Multiome data from Trevino et al (PMID: 34390642)								
3	Notes: GWAS data from Trubetskoy et al. (PMID: 35396580). Max 3 genes per peak, with 0 being strongest.								
4									
5	atac_id_PMID_34390642	rsID_PMID_35396580	position	gwas_pp_PMID_35396580	gridnet_associated_gene	variant_dist_from_TSS	gridnet_score	peak-specfic_rank	nearest_gene_to_variant
6	739	rs6673661	chr1-2441515	0.186	PLCH2	34729	0.542570487	0	PEX10
7	739	rs6673661	chr1-2441515	0.186	PEX10	28546	0.538469099	1	PEX10
8	739	rs6673661	chr1-2441515	0.186	MORN1	51064	0.533158197	2	PEX10
9	740	rs6673880	chr1-2441729	0.202	PLCH2	34119	0.541885623	0	PEX10
10	740	rs6673880	chr1-2441729	0.202	PEX10	29156	0.541201119	1	PEX10
11	740	rs6673880	chr1-2441729	0.202	MORN1	51674	0.533515977	2	PEX10
12	742	rs4648844	chr1-2443319	0.105	PLCH2	32577	0.547559269	0	PEX10
13	742	rs4648844	chr1-2443319	0.105	PEX10	30698	0.546243547	1	PEX10
14	742	rs4648844	chr1-2443319	0.105	MORN1	53216	0.538934316	2	PEX10
15	744	rs6687012	chr1-2444405	0.07	PLCH2	31423	0.546202516	0	PEX10
16	744	rs6687012	chr1-2444405	0.07	PEX10	31852	0.541857381	1	PEX10
17	744	rs6687012	chr1-2444405	0.07	MORN1	54370	0.532667588	2	PEX10
18	8248	rs12092160	chr1-28824146	0.077	TAF12	181087	0.516259	0	OPRD1
19	12113	rs11210892	chr1-43634413	0.334	KDM4A-AS1	72876	0.531369976	0	KDM4A
20	12113	rs11210892	chr1-43634413	0.334	ST3GAL3-AS1	92850	0.527466182	1	KDM4A
21	12113	rs11210892	chr1-43634413	0.334	HYI	180185	0.52307496	2	KDM4A
22	32210	rs72749142	chr1-200982179	0.215	KIF21B	23131	0.514867079	0	KIF21B
23	32220	rs12126806	chr1-200994697	0.339	KIF21B	10432	0.530049201	0	KIF21B
24	32232	rs12122721	chr1-201015352	0.114	KIF21B	9698	0.524731132	0	KIF21B

	A	B	C	D	E	F	G	H	I
25	40469	rs72759887	chr1-243457054	0.174	SDCCAG8	30562	0.515256445	0	SDCCAG8
26	41977	rs17247190	chr2-2312029	0.058	MYT1L-AS1	6893	0.578665419	0	MYT1L-AS1
27	41981	rs10206268	chr2-2322151	0.1	MYT1L-AS1	2432	0.572811384	0	MYT1L-AS1
28	41985	rs13419191	chr2-2325568	0.078	MYT1L-AS1	6159	0.579181682	0	MYT1L
29	41986	rs6708950	chr2-2326346	0.084	MYT1L-AS1	6724	0.579964542	0	MYT1L
30	41991	rs3749051	chr2-2331105	0.074	MYT1L-AS1	11414	0.575587953	0	MYT1L
31	46895	rs527882868	chr2-25216094	0.134	ARNILA	159381	0.532235201	0	LINC01381
32	46895	rs527882868	chr2-25216094	0.134	DTNB-AS1	204653	0.529724605	1	LINC01381
33	46901	rs145260949	chr2-25230278	0.375	ARNILA	145211	0.529496445	0	LINC01381
34	46901	rs145260949	chr2-25230278	0.375	DTNB-AS1	190483	0.526417128	1	LINC01381
35	46901	rs145260949	chr2-25230278	0.375	DNAJC27-AS1	228718	0.51076429	2	LINC01381
36	46904	rs114568253	chr2-25237494	0.698	ARNILA	138241	0.532404939	0	LINC01381
37	46904	rs114568253	chr2-25237494	0.698	DTNB-AS1	183513	0.527613766	1	LINC01381
38	46904	rs114568253	chr2-25237494	0.698	DNAJC27-AS1	235688	0.515999658	2	LINC01381
39	48820	rs3770752	chr2-37348993	0.289	QPCT	18989	0.575132938	0	QPCT
40	48820	rs3770752	chr2-37348993	0.289	CEBPZOS	135325	0.53993796	1	QPCT
41	48820	rs3770752	chr2-37348993	0.289	CEBPZ	117423	0.515244683	2	QPCT
42	54931	rs6722190	chr2-72941464	0.217	EMX1	8125	0.527663108	0	EMX1
43	65128	rs12991836	chr2-144383974	0.723	ZEB2-AS1	136397	0.526894864	0	GTDC1

	A	B	C	D	E	F	G	H	I
44	65128	rs12991836	chr2-144383974	0.723	GTDC1	51187	0.508338935	1	GTDC1
45	67955	rs11693702	chr2-161945674	0.061	DPP4	128238	0.531599291	0	DPP4
46	67959	rs7604885	chr2-161949898	0.065	DPP4	123809	0.533598168	0	DPP4
47	69750	rs62183854	chr2-172090705	0.098	DLX2	11880	0.574525291	0	DLX1
48	69750	rs62183854	chr2-172090705	0.098	DLX1	3153	0.573090269	1	DLX1
49	69750	rs62183854	chr2-172090705	0.098	METAP1D	19857	0.510134481	2	DLX1
50	69751	rs62183855	chr2-172091721	0.092	DLX2	10725	0.561547164	0	DLX1
51	69751	rs62183855	chr2-172091721	0.092	DLX1	4308	0.560724579	1	DLX1
52	69752	rs12469319	chr2-172092378	0.075	DLX2	10096	0.577122362	0	DLX1
53	69752	rs12469319	chr2-172092378	0.075	DLX1	4937	0.570485121	1	DLX1
54	69752	rs12469319	chr2-172092378	0.075	METAP1D	21641	0.507753565	2	DLX1
55	69768	rs62183859	chr2-172104047	0.059	DLX2	886	0.5683438	0	DLX2-DT
56	69768	rs62183859	chr2-172104047	0.059	DLX1	16419	0.556855484	1	DLX2-DT
57	69773	rs62184960	chr2-172110061	0.251	DLX2	6861	0.583559414	0	DLX2-DT
58	69773	rs62184960	chr2-172110061	0.251	DLX1	22394	0.560874212	1	DLX2-DT
59	74180	rs6736049	chr2-200350990	1	KCTD18	158842	0.509115424	0	SPATS2L
60	87494	rs4293666	chr3-30029954	0.063	RBMS3	41656	0.525636035	0	RBMS3-AS1

	A	B	C	D	E	F	G	H	I
61	90948	rs2014830	chr3-50134964	0.993	SEMA3F-AS1	20645	0.534867835	0	SEMA3F-AS1
62	90948	rs2014830	chr3-50134964	0.993	RBM5-AS1	33887	0.527057431	1	SEMA3F-AS1
63	90948	rs2014830	chr3-50134964	0.993	SLC38A3	82542	0.521447309	2	SEMA3F-AS1
64	91067	rs709210	chr3-50320438	0.609	TUSC2	7530	0.562275017	0	HYAL2
65	91067	rs709210	chr3-50320438	0.609	HYAL2	484	0.55831949	1	HYAL2
66	91067	rs709210	chr3-50320438	0.609	CYB561D2	30532	0.552351806	2	HYAL2
67	91647	rs1108842	chr3-52686064	0.076	GLT8D1	19826	0.567387528	0	GNL3
68	91647	rs1108842	chr3-52686064	0.076	NEK4	84733	0.541165287	1	GNL3
69	91647	rs1108842	chr3-52686064	0.076	SMIM4	149101	0.518793056	2	GNL3
70	93934	rs59971314	chr3-63928942	0.058	THOC7	88731	0.53377672	0	SCAANT1
71	93934	rs59971314	chr3-63928942	0.058	PSMD6-AS2	75012	0.52715185	1	SCAANT1
72	93934	rs59971314	chr3-63928942	0.058	C3orf49	105010	0.527041328	2	SCAANT1
73	110066	rs9825834	chr3-177049320	0.061	TBL1XR1-AS1	3998	0.570394935	0	TBL1XR1-AS1
74	146505	rs4132385	chr5-61319052	0.054	SMIM15-AS1	156789	0.515146482	0	ZSWIM6
75	146505	rs4132385	chr5-61319052	0.054	SMIM15	156391	0.508482321	1	ZSWIM6
76	146522	rs4604142	chr5-61346768	0.097	SMIM15-AS1	184343	0.51951195	0	ZSWIM6
77	146522	rs4604142	chr5-61346768	0.097	SMIM15	183945	0.516539263	1	ZSWIM6

	A	B	C	D	E	F	G	H	I
78	158595	rs10447226	chr5-139649225	0.056	PSD2	172484	0.511470857	0	CXXC5-AS1
79	158635	rs2336880	chr5-139684143	0.109	PSD2	137581	0.524334139	0	CXXC5-AS1
80	158636	rs10059371	chr5-139684682	0.069	PSD2	137020	0.519555187	0	CXXC5-AS1
81	158637	rs34510678	chr5-139685250	0.11	PSD2	136407	0.518737374	0	CXXC5-AS1
82	158638	rs35123781	chr5-139685595	0.074	PSD2	135878	0.51544178	0	CXXC5-AS1
83	158639	rs9687282	chr5-139686403	0.132	PSD2	135231	0.523296394	0	CXXC5-AS1
84	181358	rs76593596	chr6-96016752	0.056	FUT9	498	0.535817862	0	FUT9
85	183445	rs9374040	chr6-108676232	0.053	FOXO3	19440	0.513351999	0	LINC00222
86	189068	rs2206956	chr6-146418092	0.232	RAB32	125455	0.525876397	0	RAB32
87	189071	rs2300627	chr6-146421903	0.112	RAB32	121535	0.527661426	0	RAB32
88	196755	rs11972718	chr7-8509557	0.395	ICA1	246939	0.50544895	0	NXPH1
89	196757	rs6978582	chr7-8511018	0.097	ICA1	248372	0.512251137	0	NXPH1
90	224921	rs3808581	chr8-26392531	0.146	BNIP3L	2729	0.540524783	0	BNIP3L
91	225237	rs2472554	chr8-27470837	0.134	EPHX2	44477	0.560356245	0	CHRNA2
92	236502	rs4734654	chr8-102657763	0.154	KLF10	3121	0.571043948	0	KLF10
93	243774	rs10092551	chr8-143793775	0.166	IQANK1	4146	0.572068173	0	SCRIB
94	243774	rs10092551	chr8-143793775	0.166	SCRIB	14516	0.566340553	1	SCRIB
95	243774	rs10092551	chr8-143793775	0.166	ZNF707	85610	0.5282787	2	SCRIB

	A	B	C	D	E	F	G	H	I
96	243775	rs11993154	chr8-143794599	0.059	IQANK1	5022	0.566030711	0	SCRIB
97	243775	rs11993154	chr8-143794599	0.059	SCRIB	13640	0.555824944	1	SCRIB
98	243775	rs11993154	chr8-143794599	0.059	ZNF707	86486	0.517503918	2	SCRIB
99	243799	rs55930529	chr8-143823571	0.155	IQANK1	33825	0.563010124	0	PUF60
100	243799	rs55930529	chr8-143823571	0.155	SCRIB	14663	0.551028094	1	PUF60
101	243799	rs55930529	chr8-143823571	0.155	NRBP2	16790	0.526679617	2	PUF60
102	249506	rs9644880	chr9-36318848	0.249	GNE	41592	0.548573002	0	GNE
103	251091	rs500102	chr9-74743829	0.271	RORB-AS1	244279	0.533410233	0	TRPM6
104	251091	rs500102	chr9-74743829	0.271	TRPM6	143814	0.516731714	1	TRPM6
105	251091	rs500102	chr9-74743829	0.271	C9orf40	209006	0.508641888	2	TRPM6
106	254204	rs3815563	chr9-93450864	0.076	FAM120A OS	148	0.563113902	0	FAM120A
107	254204	rs3815563	chr9-93450864	0.076	C9orf129	104479	0.533813833	1	FAM120A
108	254204	rs3815563	chr9-93450864	0.076	PHF2	125496	0.506014514	2	FAM120A
109	254219	rs7048853	chr9-93490896	0.123	C9orf129	144475	0.531655177	0	FAM120A OS
110	254219	rs7048853	chr9-93490896	0.123	FAM120A OS	39644	0.530378251	1	FAM120A OS
111	254219	rs7048853	chr9-93490896	0.123	PHF2	85500	0.507471776	2	FAM120A OS
112	254223	rs59582399	chr9-93523881	0.072	FAM120A OS	72777	0.521144356	0	PHF2

	A	B	C	D	E	F	G	H	I
113	254223	rs59582399	chr9-93523881	0.072	C9orf129	177608	0.519605621	1	PHF2
114	254223	rs59582399	chr9-93523881	0.072	PHF2	52367	0.513118383	2	PHF2
115	262121	rs77206190	chr9-132017990	0.052	NTNG2	143463	0.524969583	0	MED27
116	281589	rs10883761	chr10-102660261	0.052	SFXN2	65842	0.531024866	0	TRIM8
117	281589	rs10883761	chr10-102660261	0.052	WBP1L	172063	0.528737152	1	TRIM8
118	281589	rs10883761	chr10-102660261	0.052	C10orf95-AS1	206318	0.516344221	2	TRIM8
119	281592	rs4244354	chr10-102666420	0.061	SFXN2	60003	0.527725981	0	TRIM8
120	281592	rs4244354	chr10-102666420	0.061	WBP1L	166224	0.521215382	1	TRIM8
121	281592	rs4244354	chr10-102666420	0.061	C10orf95-AS1	212157	0.512767191	2	TRIM8
122	299026	rs58950470	chr11-65616284	0.167	MAP3K11	2012	0.562014346	0	PCNX3
123	299026	rs58950470	chr11-65616284	0.167	PCNX3	13663	0.555640335	1	PCNX3
124	299026	rs58950470	chr11-65616284	0.167	RNASEH2C	104050	0.544327436	2	PCNX3
125	307188	rs7112715	chr11-113532053	0.499	TMPRSS5	174149	0.522753169	0	DRD2
126	307188	rs7112715	chr11-113532053	0.499	TTC12	215485	0.510079731	1	DRD2
127	307201	rs7928017	chr11-113578040	0.146	TMPRSS5	127878	0.518010718	0	DRD2
128	307201	rs7928017	chr11-113578040	0.146	ZW10	195240	0.513555678	1	DRD2
129	310377	rs7927437	chr11-123525279	0.134	ZNF202	216371	0.514258875	0	GRAMD1B

	A	B	C	D	E	F	G	H	I
130	312920	rs502834	chr11-133947438	0.08	IGSF9B	8990	0.543228502	0	IGSF9B
131	312949	rs4936216	chr11-133983113	0.097	IGSF9B	26101	0.529105423	0	IGSF9B
132	312971	rs78971498	chr11-134014628	0.108	IGSF9B	57698	0.514349952	0	LINC02730
133	312971	rs78971498	chr11-134014628	0.108	ACAD8	247233	0.505918393	1	LINC02730
134	312971	rs78971498	chr11-134014628	0.108	THYN1	238349	0.505590542	2	LINC02730
135	312974	rs58406308	chr11-134023085	0.07	THYN1	230134	0.519307004	0	LINC02730
136	312974	rs58406308	chr11-134023085	0.07	ACAD8	239018	0.516656094	1	LINC02730
137	312974	rs58406308	chr11-134023085	0.07	IGSF9B	65913	0.515617228	2	LINC02730
138	312980	rs7938834	chr11-134029834	0.055	IGSF9B	72631	0.508912703	0	LINC02730
139	312980	rs7938834	chr11-134029834	0.055	THYN1	223416	0.505300548	1	LINC02730
140	312985	rs2236657	chr11-134034290	0.073	THYN1	218901	0.516231899	0	LINC02731
141	312985	rs2236657	chr11-134034290	0.073	IGSF9B	77146	0.512657549	1	LINC02731
142	312985	rs2236657	chr11-134034290	0.073	ACAD8	227785	0.511137995	2	LINC02731
143	312987	rs687187	chr11-134035439	0.083	IGSF9B	78635	0.51213344	0	LINC02731
144	312987	rs687187	chr11-134035439	0.083	THYN1	217412	0.509856118	1	LINC02731
145	313082	rs11223774	chr11-134377421	0.394	GLB1L2	45508	0.540910206	0	B3GAT1
146	313082	rs11223774	chr11-134377421	0.394	THYN1	124081	0.521707493	1	B3GAT1

	A	B	C	D	E	F	G	H	I
147	313082	rs11223774	chr11-134377421	0.394	ACAD8	115197	0.52013415	2	B3GAT1
148	313823	rs11062170	chr12-2239678	0.095	DCP1B	234981	0.516765514	0	CACNA1C-AS4
149	320649	rs706790	chr12-50070687	0.07	ASIC1	9584	0.539298286	0	ASIC1
150	320649	rs706790	chr12-50070687	0.07	COX14	41323	0.523805669	1	ASIC1
151	320649	rs706790	chr12-50070687	0.07	NCKAP5L	241664	0.523190087	2	ASIC1
152	322233	rs324017	chr12-57094031	0.962	NAB2	1695	0.561457227	0	NAB2
153	322233	rs324017	chr12-57094031	0.962	STAT6	17069	0.557972495	1	NAB2
154	322233	rs324017	chr12-57094031	0.962	NXPH4	125525	0.51878358	2	NAB2
155	329143	rs4246263	chr12-104243799	0.054	EID3	59709	0.54544124	0	TXNRD1
156	329143	rs4246263	chr12-104243799	0.054	NFYB	105289	0.520740127	1	TXNRD1
157	329143	rs4246263	chr12-104243799	0.054	GLT8D2	193394	0.514301973	2	TXNRD1
158	332744	rs77278225	chr12-120844458	0.287	C12orf43	171790	0.511160103	0	SPPL3
159	332751	rs7486605	chr12-120903362	0.981	C12orf43	112918	0.51529939	0	SPPL3
160	332751	rs7486605	chr12-120903362	0.981	P2RX7	229409	0.509123176	1	SPPL3
161	333161	rs1979237	chr12-122841348	1	CCDC62	52548	0.530261305	0	HIP1R
162	340007	rs953102	chr13-43699249	0.1	CCDC122	180095	0.521411178	0	ENOX1
163	364705	rs11847162	chr14-103346720	0.112	EIF5	5489	0.50563941	0	EIF5

	A	B	C	D	E	F	G	H	I
164	367234	rs117799466	chr15-34367316	0.906	SLC12A6	29263	0.545975566	0	LPCAT4
165	367234	rs117799466	chr15-34367316	0.906	LPCAT4	24	0.537373474	1	LPCAT4
166	367234	rs117799466	chr15-34367316	0.906	NOP10	23869	0.53358522	2	LPCAT4
167	368375	rs56205728	chr15-40275036	0.543	PAK6	692	0.579155222	0	ANKRD63
168	368375	rs56205728	chr15-40275036	0.543	INAFM2	50094	0.537601762	1	ANKRD63
169	368375	rs56205728	chr15-40275036	0.543	BMF	165744	0.524451257	2	ANKRD63
170	377575	rs7178152	chr15-89243824	0.152	RLBP1	22070	0.541695045	0	FANCI
171	377660	rs208827	chr15-89399100	0.066	LINC00928	124676	0.529049087	0	MIR9-3HG
172	377660	rs208827	chr15-89399100	0.066	RLBP1	177268	0.527834436	1	MIR9-3HG
173	377660	rs208827	chr15-89399100	0.066	POLG	64429	0.526208416	2	MIR9-3HG
174	377662	rs182317	chr15-89400370	0.055	POLG	65737	0.530252145	0	MIR9-3HG
175	377662	rs182317	chr15-89400370	0.055	RLBP1	178576	0.521037796	1	MIR9-3HG
176	377662	rs182317	chr15-89400370	0.055	LINC00928	123368	0.513989729	2	MIR9-3HG
177	378059	rs4702	chr15-90883330	0.999	MAN2A2	20408	0.571502915	0	FES
178	378059	rs4702	chr15-90883330	0.999	FURIN	1867	0.558483091	1	FES
179	378059	rs4702	chr15-90883330	0.999	UNC45A	46398	0.553085894	2	FES
180	396302	rs4790338	chr17-1371519	0.095	INPP5K	144916	0.518666926	0	YWHAE
181	396302	rs4790338	chr17-1371519	0.095	MYO1C	119492	0.5175162	1	YWHAE

	A	B	C	D	E	F	G	H	I
182	396302	rs4790338	chr17-1371519	0.095	PITPNA-AS1	145186	0.507254385	2	YWHAE
183	396316	rs28365859	chr17-1400484	0.165	INPP5K	116101	0.534642969	0	YWHAE
184	396316	rs28365859	chr17-1400484	0.165	MYO1C	90677	0.532800396	1	YWHAE
185	396316	rs28365859	chr17-1400484	0.165	PITPNA-AS1	116371	0.519511352	2	YWHAE
186	398790	rs9908102	chr17-12993236	0.067	ELAC2	24600	0.535925249	0	ELAC2
187	399903	rs959071	chr17-19238913	0.382	EPN2-AS1	67032	0.518669111	0	EPN2
188	399903	rs959071	chr17-19238913	0.382	B9D1	138964	0.507006818	1	EPN2
189	399947	rs12602286	chr17-19333641	0.172	MFAP4	53062	0.553732851	0	EPN2-AS1
190	399947	rs12602286	chr17-19333641	0.172	EPN2-AS1	27367	0.546348867	1	EPN2-AS1
191	399947	rs12602286	chr17-19333641	0.172	B9D1	44065	0.538736664	2	EPN2-AS1
192	407176	rs34888090	chr17-57664675	0.088	MSI2	10078	0.507750863	0	CCDC182
193	407178	rs12938078	chr17-57665916	0.073	MRPS23	183806	0.515961292	0	CCDC182
194	407178	rs12938078	chr17-57665916	0.073	MSI2	8781	0.513769988	1	CCDC182
195	408133	rs4291	chr17-63476833	0.087	CYB561	35590	0.547801473	0	ACE
196	408133	rs4291	chr17-63476833	0.087	LIMD2	224243	0.523080934	1	ACE
197	408133	rs4291	chr17-63476833	0.087	TACO1	124646	0.519911228	2	ACE
198	408136	rs4295	chr17-63478937	0.053	CYB561	37984	0.543418646	0	ACE
199	408136	rs4295	chr17-63478937	0.053	TACO1	122252	0.517891458	1	ACE

	A	B	C	D	E	F	G	H	I
200	408136	rs4295	chr17-63478937	0.053	LIMD2	221849	0.516083783	2	ACE
201	422428	rs4128242	chr18-55080458	0.095	RAB27B	195621	0.511293109	0	LINC01929
202	422429	rs12969453	chr18-55084477	0.121	RAB27B	199757	0.518613879	0	LINC01929
203	422561	rs74914300	chr18-55343117	0.058	TCF4-AS2	148988	0.526100813	0	TCF4-AS1
204	422592	rs75037433	chr18-55390864	0.056	TCF4-AS2	100863	0.532477931	0	TCF4-AS1
205	422592	rs75037433	chr18-55390864	0.056	TCF4-AS1	61221	0.510827016	1	TCF4-AS1
206	422598	rs78322266	chr18-55396445	0.064	TCF4-AS2	95420	0.532373459	0	TCF4-AS1
207	422598	rs78322266	chr18-55396445	0.064	TCF4-AS1	55778	0.513108253	1	TCF4-AS1
208	426658	rs509573	chr18-79800879	0.107	KCNG2	74619	0.53759408	0	KCNG2
209	426658	rs509573	chr18-79800879	0.107	SLC66A2	150392	0.519656695	1	KCNG2
210	426658	rs509573	chr18-79800879	0.107	CTDP1	64412	0.506286845	2	KCNG2
211	426663	rs4798923	chr18-79869373	0.07	KCNG2	6133	0.570059436	0	KCNG2
212	426663	rs4798923	chr18-79869373	0.07	HSBP1L1	100292	0.533798604	1	KCNG2
213	426663	rs4798923	chr18-79869373	0.07	SLC66A2	81906	0.528829274	2	KCNG2
214	426664	rs11660941	chr18-79870600	0.07	KCNG2	5051	0.578060168	0	KCNG2
215	426664	rs11660941	chr18-79870600	0.07	SLC66A2	80824	0.537551811	1	KCNG2
216	426664	rs11660941	chr18-79870600	0.07	HSBP1L1	99210	0.533140189	2	KCNG2

	A	B	C	D	E	F	G	H	I
217	426665	rs61090726	chr18-79871219	0.063	KCNG2	4367	0.575771059	0	KCNG2
218	426665	rs61090726	chr18-79871219	0.063	SLC66A2	80140	0.537748608	1	KCNG2
219	426665	rs61090726	chr18-79871219	0.063	HSBP1L1	98526	0.537262121	2	KCNG2
220	426666	rs72980085	chr18-79871679	0.08	KCNG2	3769	0.577368456	0	KCNG2
221	426666	rs72980085	chr18-79871679	0.08	SLC66A2	79542	0.545208017	1	KCNG2
222	426666	rs72980085	chr18-79871679	0.08	HSBP1L1	97928	0.533410957	2	KCNG2
223	426667	rs8091497	chr18-79872565	0.09	KCNG2	2873	0.560140138	0	KCNG2
224	426667	rs8091497	chr18-79872565	0.09	SLC66A2	78646	0.538529109	1	KCNG2
225	426667	rs8091497	chr18-79872565	0.09	HSBP1L1	97032	0.518167152	2	KCNG2
226	430431	rs322124	chr19-11288696	0.239	DOCK6	34172	0.554087633	0	TSPAN16
227	430431	rs322124	chr19-11288696	0.239	TMEM205	57542	0.550831696	1	TSPAN16
228	430431	rs322124	chr19-11288696	0.239	CCDC159	63242	0.548895192	2	TSPAN16
229	430433	rs322129	chr19-11290429	0.223	DOCK6	36117	0.543044544	0	TSPAN16
230	430433	rs322129	chr19-11290429	0.223	CCDC159	61297	0.541698906	1	TSPAN16
231	430433	rs322129	chr19-11290429	0.223	TMEM205	55597	0.535393988	2	TSPAN16
232	430560	rs72986630	chr19-11738921	0.99	ZNF823	247	0.559184299	0	ZNF823
233	430560	rs72986630	chr19-11738921	0.99	ZNF441	27744	0.552197366	1	ZNF823

	A	B	C	D	E	F	G	H	I
234	430560	rs72986630	chr19-11738921	0.99	ZNF491	59332	0.534652146	2	ZNF823
235	437940	rs2304204	chr19-49665763	0.09	SCAF1	20729	0.558311257	0	IRF3
236	437940	rs2304204	chr19-49665763	0.09	IRF3	102	0.548925285	1	IRF3
237	437940	rs2304204	chr19-49665763	0.09	CPT1C	46841	0.547245293	2	IRF3
238	439190	rs758749	chr19-56678350	0.346	ZNF71	82771	0.547101722	0	ZNF835
239	439190	rs758749	chr19-56678350	0.346	MIMT1	162643	0.540521334	1	ZNF835
240	439190	rs758749	chr19-56678350	0.346	ZNF835	6323	0.538772949	2	ZNF835

	A	B	C	D	E	F
1	**Table S2**					
2	gene_from_gridnet	tf	log10VariantEffect_from_FIMO	position	rsid_from_PMID_35396580	effect_on_binding_affinity
3	TBL1XR1-AS1	KLF15	-4.07849	chr3-176767108	rs9825834	TF loss
4	TBL1XR1-AS1	SP5	-3.89992	chr3-176767108	rs9825834	TF loss
5	TBL1XR1-AS1	MAZ	-3.69357	chr3-176767108	rs9825834	TF loss
6	TBL1XR1-AS1	MAZ	-3.31091	chr3-176767108	rs9825834	TF loss
7	TBL1XR1-AS1	SP1	-2.72236	chr3-176767108	rs9825834	TF loss
8	MAP3K11, PCNX3, RNASEH2C	KLF6	2.640419	chr11-65383755	rs58950470	TF gain
9	TBL1XR1-AS1	ZN263	2.609962	chr3-176767108	rs9825834	TF gain
10	MAP3K11, PCNX3, RNASEH2C	PLAG1	-2.5616	chr11-65383755	rs58950470	TF loss
11	TBL1XR1-AS1	KLF15	-2.55654	chr3-176767108	rs9825834	TF loss
12	SLC12A6, LPCAT4, NOP10	KLF6	-2.52533	chr15-34659517	rs117799466	TF loss
13	TBL1XR1-AS1	VEZF1	-2.51755	chr3-176767108	rs9825834	TF loss
14	TBL1XR1-AS1	KLF1	-2.40514	chr3-176767108	rs9825834	TF loss
15	TBL1XR1-AS1	SP1	-2.40302	chr3-176767108	rs9825834	TF loss

	A	B	C	D	E	F
16	NAB2, STAT6, NXPH4	PURA	-2.39794	chr12-57487814	rs324017	TF loss
17	TBL1XR1-AS1	KLF5	-2.38352	chr3-176767108	rs9825834	TF loss
18	SLC12A6, LPCAT4, NOP10	SP1	2.373195	chr15-34659517	rs117799466	TF gain
19	TBL1XR1-AS1	SP3	-2.35482	chr3-176767108	rs9825834	TF loss
20	NAB2, STAT6, NXPH4	PURA	-2.33488	chr12-57487814	rs324017	TF loss
21	SLC12A6, LPCAT4, NOP10	ZN281	2.232606	chr15-34659517	rs117799466	TF gain
22	MAP3K11, PCNX3, RNASEH2C	WT1	-2.2213	chr11-65383755	rs58950470	TF loss
23	MAP3K11, PCNX3, RNASEH2C	SP1	-2.21927	chr11-65383755	rs58950470	TF loss
24	MAP3K11, PCNX3, RNASEH2C	SP3	-2.2165	chr11-65383755	rs58950470	TF loss
25	TBL1XR1-AS1	ZNF281	-2.19735	chr3-176767108	rs9825834	TF loss
26	TBL1XR1-AS1	KLF16	-2.1916	chr3-176767108	rs9825834	TF loss
27	SLC12A6, LPCAT4, NOP10	MAZ	2.183573	chr15-34659517	rs117799466	TF gain
28	TBL1XR1-AS1	PATZ1	-2.10984	chr3-176767108	rs9825834	TF loss
29	ARNILA, DTNB-AS1	KLF6	-2.10134	chr2-25438963	rs527882868	TF loss
30	SLC12A6, LPCAT4, NOP10	SP2	2.095089	chr15-34659517	rs117799466	TF gain

	A	B	C	D	E	F
31	MAP3K11, PCNX3, RNASEH2C	PATZ1	-2.07479	chr11-65383755	rs58950470	TF loss
32	ICA1	ZNF440	2.074119	chr7-8549187	rs11972718	TF gain
33	SLC12A6, LPCAT4, NOP10	HKR1	2.051694	chr15-34659517	rs117799466	TF gain
34	TBL1XR1-AS1	SP2	-2.01878	chr3-176767108	rs9825834	TF loss
35	EPHX2	KLF6	-2.00421	chr8-27328354	rs2472554	TF loss
36	SLC12A6, LPCAT4, NOP10	SP4	2.002719	chr15-34659517	rs117799466	TF gain
37	TBL1XR1-AS1	KLF1	-1.98038	chr3-176767108	rs9825834	TF loss
38	TBL1XR1-AS1	KLF15	1.978058	chr3-176767108	rs9825834	TF gain
39	TBL1XR1-AS1	KLF6	-1.9778	chr3-176767108	rs9825834	TF loss
40	MAP3K11, PCNX3, RNASEH2C	ZNF37A	-1.97126	chr11-65383755	rs58950470	TF loss
41	MAP3K11, PCNX3, RNASEH2C	SP3	-1.96586	chr11-65383755	rs58950470	TF loss
42	SLC12A6, LPCAT4, NOP10	OSR2	-1.95823	chr15-34659517	rs117799466	TF loss
43	SLC12A6, LPCAT4, NOP10	ZN281	1.896833	chr15-34659517	rs117799466	TF gain
44	SLC12A6, LPCAT4, NOP10	SP4	1.886301	chr15-34659517	rs117799466	TF gain
45	MAP3K11, PCNX3, RNASEH2C	MAZ	-1.87932	chr11-65383755	rs58950470	TF loss

	A	B	C	D	E	F
46	SLC12A6, LPCAT4, NOP10	SP1	1.875827	chr15-34659517	rs117799466	TF gain
47	TBL1XR1-AS1	KLF12	-1.86923	chr3-176767108	rs9825834	TF loss
48	SLC12A6, LPCAT4, NOP10	ZSC22	1.867134	chr15-34659517	rs117799466	TF gain
49	MAP3K11, PCNX3, RNASEH2C	ZNF707	1.848676	chr11-65383755	rs58950470	TF gain
50	EMX1	VEZF1	-1.84253	chr2-73168593	rs6722190	TF loss
51	TBL1XR1-AS1	MAZ	1.84018	chr3-176767108	rs9825834	TF gain
52	SLC12A6, LPCAT4, NOP10	WT1	1.830145	chr15-34659517	rs117799466	TF gain
53	MAP3K11, PCNX3, RNASEH2C	ZNF148	-1.82925	chr11-65383755	rs58950470	TF loss
54	MAP3K11, PCNX3, RNASEH2C	SP1	-1.82729	chr11-65383755	rs58950470	TF loss
55	SLC12A6, LPCAT4, NOP10	VEZF1	1.826595	chr15-34659517	rs117799466	TF gain
56	MAP3K11, PCNX3, RNASEH2C	SP3	-1.81299	chr11-65383755	rs58950470	TF loss
57	MAP3K11, PCNX3, RNASEH2C	TBX15	-1.80695	chr11-65383755	rs58950470	TF loss
58	SLC12A6, LPCAT4, NOP10	ZNF148	1.793663	chr15-34659517	rs117799466	TF gain
59	SLC12A6, LPCAT4, NOP10	ZBT17	1.785063	chr15-34659517	rs117799466	TF gain
60	SLC12A6, LPCAT4, NOP10	PATZ1	1.774062	chr15-34659517	rs117799466	TF gain

	A	B	C	D	E	F
61	NAB2, STAT6, NXPH4	ZNF780A	-1.76737	chr12-57487814	rs324017	TF loss
62	NAB2, STAT6, NXPH4	ZN335	-1.75868	chr12-57487814	rs324017	TF loss
63	MAP3K11, PCNX3, RNASEH2C	KLF3	1.752558	chr11-65383755	rs58950470	TF gain
64	SLC12A6, LPCAT4, NOP10	PRDM9	1.727846	chr15-34659517	rs117799466	TF gain
65	MAP3K11, PCNX3, RNASEH2C	MAZ	-1.7275	chr11-65383755	rs58950470	TF loss
66	TBL1XR1-AS1	WT1	-1.72163	chr3-176767108	rs9825834	TF loss
67	TBL1XR1-AS1	ZN467	-1.72141	chr3-176767108	rs9825834	TF loss
68	TBL1XR1-AS1	ZNF281	-1.72127	chr3-176767108	rs9825834	TF loss
69	SLC12A6, LPCAT4, NOP10	SP4	-1.71551	chr15-34659517	rs117799466	TF loss
70	KLF10	SP4	-1.71018	chr8-103669991	rs4734654	TF loss
71	DLX2, DLX1	SP2	1.708357	chr2-172968775	rs62183859	TF gain
72	MAP3K11, PCNX3, RNASEH2C	SP1	-1.70466	chr11-65383755	rs58950470	TF loss
73	MAP3K11, PCNX3, RNASEH2C	SP2	-1.69734	chr11-65383755	rs58950470	TF loss
74	SLC12A6, LPCAT4, NOP10	ZNF281	1.697203	chr15-34659517	rs117799466	TF gain
75	C12orf43, P2RX7	ZNF610	-1.68518	chr12-121341165	rs7486605	TF loss

	A	B	C	D	E	F
76	SFXN2, WBP1L, C10orf95-AS1	ZN713	1.670559	chr10-104420018	rs10883761	TF gain
77	ARNILA, DTNB-AS1	KLF1	1.66549	chr2-25438963	rs527882868	TF gain
78	MAP3K11, PCNX3, RNASEH2C	ZN148	-1.65392	chr11-65383755	rs58950470	TF loss
79	SLC12A6, LPCAT4, NOP10	SP3	1.651109	chr15-34659517	rs117799466	TF gain
80	SLC12A6, LPCAT4, NOP10	SP2	1.65042	chr15-34659517	rs117799466	TF gain
81	MAP3K11, PCNX3, RNASEH2C	MAZ	-1.6503	chr11-65383755	rs58950470	TF loss
82	MAP3K11, PCNX3, RNASEH2C	ZN263	-1.64724	chr11-65383755	rs58950470	TF loss
83	MFAP4, EPN2-AS1, B9D1	ZNF267	-1.63259	chr17-19236954	rs12602286	TF loss
84	MAP3K11, PCNX3, RNASEH2C	MAZ	-1.62598	chr11-65383755	rs58950470	TF loss
85	EPHX2	KLF16	1.619689	chr8-27328354	rs2472554	TF gain
86	DLX2, DLX1	ZFX	-1.61621	chr2-172974789	rs62184960	TF loss
87	SLC12A6, LPCAT4, NOP10	SP3	1.606717	chr15-34659517	rs117799466	TF gain
88	SFXN2, WBP1L, C10orf95-AS1	ZNF823	1.606381	chr10-104420018	rs10883761	TF gain
89	MAP3K11, PCNX3, RNASEH2C	ZNF460	-1.60594	chr11-65383755	rs58950470	TF loss
90	TBL1XR1-AS1	ZNF148	-1.60128	chr3-176767108	rs9825834	TF loss

	A	B	C	D	E	F
91	KCNG2, SLC66A2, HSBP1L1	SP1	-1.60085	chr18-77632565	rs8091497	TF loss
92	SLC12A6, LPCAT4, NOP10	VEZF1	1.59936	chr15-34659517	rs117799466	TF gain
93	MAP3K11, PCNX3, RNASEH2C	KLF3	-1.59146	chr11-65383755	rs58950470	TF loss
94	TBL1XR1-AS1	KLF6	1.589185	chr3-176767108	rs9825834	TF gain
95	ZNF71, MIMT1, ZNF835	ZNF610	-1.58893	chr19-57189718	rs758749	TF loss
96	TBL1XR1-AS1	VEZF1	1.588413	chr3-176767108	rs9825834	TF gain
97	TBL1XR1-AS1	ZN281	-1.58827	chr3-176767108	rs9825834	TF loss
98	TBL1XR1-AS1	ZNF398	1.585789	chr3-176767108	rs9825834	TF gain
99	MAP3K11, PCNX3, RNASEH2C	ZNF467	-1.58246	chr11-65383755	rs58950470	TF loss
100	SLC12A6, LPCAT4, NOP10	EGR2	1.567812	chr15-34659517	rs117799466	TF gain
101	SLC12A6, LPCAT4, NOP10	ZNF281	1.548375	chr15-34659517	rs117799466	TF gain
102	SLC12A6, LPCAT4, NOP10	ZNF460	1.539133	chr15-34659517	rs117799466	TF gain
103	DLX2, DLX1	SP3	1.53844	chr2-172968775	rs62183859	TF gain
104	TBL1XR1-AS1	ZN148	-1.53494	chr3-176767108	rs9825834	TF loss
105	SLC12A6, LPCAT4, NOP10	EGR1	1.527402	chr15-34659517	rs117799466	TF gain

	A	B	C	D	E	F
106	DLX2, DLX1	KLF15	-1.51964	chr2-172968775	rs62183859	TF loss
107	KLF10	TBX1	1.511588	chr8-103669991	rs4734654	TF gain
108	MAP3K11, PCNX3, RNASEH2C	KLF15	-1.50922	chr11-65383755	rs58950470	TF loss
109	NAB2, STAT6, NXPH4	ZN335	-1.50775	chr12-57487814	rs324017	TF loss
110	TBL1XR1-AS1	MAZ	1.505464	chr3-176767108	rs9825834	TF gain
111	DLX2, DLX1	KLF16	1.500165	chr2-172968775	rs62183859	TF gain
112	ARNILA, DTNB-AS1	SP1	1.499813	chr2-25438963	rs527882868	TF gain
113	TBL1XR1-AS1	MAZ	-1.49971	chr3-176767108	rs9825834	TF loss
114	PLCH2, PEX10, MORN1	SP3	1.499231	chr1-2374758	rs4648844	TF gain
115	MAP3K11, PCNX3, RNASEH2C	TBX15	-1.49917	chr11-65383755	rs58950470	TF loss
116	KLF10	KLF15	-1.49714	chr8-103669991	rs4734654	TF loss
117	TBL1XR1-AS1	MAZ	-1.49334	chr3-176767108	rs9825834	TF loss
118	KCNG2, SLC66A2, HSBP1L1	ZN770	-1.48734	chr18-77631679	rs72980085	TF loss
119	SLC12A6, LPCAT4, NOP10	Wt1	1.486498	chr15-34659517	rs117799466	TF gain
120	SLC12A6, LPCAT4, NOP10	SP2	-1.48578	chr15-34659517	rs117799466	TF loss

	A	B	C	D	E	F
121	SLC12A6, LPCAT4, NOP10	WT1	1.483945	chr15-34659517	rs117799466	TF gain
122	TBL1XR1-AS1	ZNF468	-1.4829	chr3-176767108	rs9825834	TF loss
123	ZNF202	ZNF341	-1.47893	chr11-123395987	rs7927437	TF loss
124	PLCH2, PEX10, MORN1	VEZF1	-1.47815	chr1-2374758	rs4648844	TF loss
125	TBL1XR1-AS1	ZNF148	1.47814	chr3-176767108	rs9825834	TF gain
126	MAP3K11, PCNX3, RNASEH2C	E2F1	-1.47799	chr11-65383755	rs58950470	TF loss
127	EPHX2	ZNF132	1.475882	chr8-27328354	rs2472554	TF gain
128	SLC12A6, LPCAT4, NOP10	TBX15	1.475587	chr15-34659517	rs117799466	TF gain
129	MYT1L-AS1	TBX15	1.473191	chr2-2330118	rs6708950	TF gain
130	PLCH2, PEX10, MORN1	SP2	1.472712	chr1-2374758	rs4648844	TF gain
131	NAB2, STAT6, NXPH4	ZN770	-1.4719	chr12-57487814	rs324017	TF loss
132	MAP3K11, PCNX3, RNASEH2C	KLF12	-1.46699	chr11-65383755	rs58950470	TF loss
133	MAP3K11, PCNX3, RNASEH2C	ZNF460	-1.46152	chr11-65383755	rs58950470	TF loss
134	SLC12A6, LPCAT4, NOP10	MAZ	1.461301	chr15-34659517	rs117799466	TF gain
135	TBL1XR1-AS1	TBX15	-1.45976	chr3-176767108	rs9825834	TF loss
136	IGSF9B, THYN1	ZNF282	1.459585	chr11-133905334	rs687187	TF gain

	A	B	C	D	E	F
137	SLC12A6, LPCAT4, NOP10	EGR2	1.459077	chr15-34659517	rs117799466	TF gain
138	SLC12A6, LPCAT4, NOP10	ZNF774	1.457779	chr15-34659517	rs117799466	TF gain
139	DLX2, DLX1	ZNF180	1.455626	chr2-172968775	rs62183859	TF gain
140	SLC12A6, LPCAT4, NOP10	SP1	1.450544	chr15-34659517	rs117799466	TF gain
141	ARNILA, DTNB-AS1	ZNF534	1.449192	chr2-25438963	rs527882868	TF gain
142	MAP3K11, PCNX3, RNASEH2C	ZNF341	-1.44909	chr11-65383755	rs58950470	TF loss
143	TBL1XR1-AS1	ZN263	-1.44768	chr3-176767108	rs9825834	TF loss
144	ARNILA, DTNB-AS1	KLF3	-1.44569	chr2-25438963	rs527882868	TF loss
145	ASIC1, COX14, NCKAP5L	ZNF320	1.44364	chr12-50464470	rs706790	TF gain
146	MAP3K11, PCNX3, RNASEH2C	ZNF180	-1.44305	chr11-65383755	rs58950470	TF loss
147	MAP3K11, PCNX3, RNASEH2C	E2F1	-1.44267	chr11-65383755	rs58950470	TF loss
148	KLF10	VEZF1	-1.4415	chr8-103669991	rs4734654	TF loss
149	MAP3K11, PCNX3, RNASEH2C	ZNF468	-1.43806	chr11-65383755	rs58950470	TF loss
150	SLC12A6, LPCAT4, NOP10	EGR2	1.436503	chr15-34659517	rs117799466	TF gain
151	SLC12A6, LPCAT4, NOP10	ZNF281	1.435423	chr15-34659517	rs117799466	TF gain

	A	B	C	D	E	F
152	MRPS23, MSI2	ZNF580	-1.43537	chr17-55743277	rs12938078	TF loss
153	SLC12A6, LPCAT4, NOP10	EGR2	1.425051	chr15-34659517	rs117799466	TF gain
154	TBL1XR1-AS1	ZNF467	-1.41899	chr3-176767108	rs9825834	TF loss
155	NAB2, STAT6, NXPH4	PRDM9	-1.41452	chr12-57487814	rs324017	TF loss
156	EMX1	MAZ	-1.41173	chr2-73168593	rs6722190	TF loss
157	TBL1XR1-AS1	ZNF467	1.40981	chr3-176767108	rs9825834	TF gain
158	MAP3K11, PCNX3, RNASEH2C	ZN148	-1.40902	chr11-65383755	rs58950470	TF loss
159	SLC12A6, LPCAT4, NOP10	KLF15	1.406484	chr15-34659517	rs117799466	TF gain
160	TBL1XR1-AS1	ZN467	-1.39717	chr3-176767108	rs9825834	TF loss
161	TBL1XR1-AS1	SP4	-1.39332	chr3-176767108	rs9825834	TF loss
162	KLF10	ZBT17	-1.39222	chr8-103669991	rs4734654	TF loss
163	ARNILA, DTNB-AS1	SP1	1.391814	chr2-25438963	rs527882868	TF gain
164	PLCH2, PEX10, MORN1	TBX1	1.391458	chr1-2374758	rs4648844	TF gain
165	TBL1XR1-AS1	ZN467	1.38489	chr3-176767108	rs9825834	TF gain
166	SLC12A6, LPCAT4, NOP10	SP2	1.384542	chr15-34659517	rs117799466	TF gain

	A	B	C	D	E	F
167	TAF12	ZNF496	-1.38102	chr1-29150658	rs12092160	TF loss
168	TBL1XR1-AS1	SP5	1.380485	chr3-176767108	rs9825834	TF gain
169	MAP3K11, PCNX3, RNASEH2C	PLAG1	-1.37827	chr11-65383755	rs58950470	TF loss
170	NAB2, STAT6, NXPH4	HKR1	-1.37727	chr12-57487814	rs324017	TF loss
171	KCNG2, SLC66A2, CTDP1	ZN341	-1.37634	chr18-77560879	rs509573	TF loss
172	MAP3K11, PCNX3, RNASEH2C	PATZ1	-1.37358	chr11-65383755	rs58950470	TF loss
173	SLC12A6, LPCAT4, NOP10	SP2	1.370303	chr15-34659517	rs117799466	TF gain
174	PLCH2, PEX10, MORN1	PURA	1.367851	chr1-2374758	rs4648844	TF gain
175	TBL1XR1-AS1	ZNF444	-1.3637	chr3-176767108	rs9825834	TF loss
176	MAP3K11, PCNX3, RNASEH2C	CTCFL	-1.36228	chr11-65383755	rs58950470	TF loss
177	MRPS23, MSI2	KLF16	-1.3614	chr17-55743277	rs12938078	TF loss
178	TBL1XR1-AS1	KLF1	1.361376	chr3-176767108	rs9825834	TF gain
179	PSD2	ZNF571	-1.36103	chr5-139028810	rs10447226	TF loss
180	NAB2, STAT6, NXPH4	KLF3	-1.36057	chr12-57487814	rs324017	TF loss
181	SLC12A6, LPCAT4, NOP10	SP3	1.359506	chr15-34659517	rs117799466	TF gain

	A	B	C	D	E	F
182	MAP3K11, PCNX3, RNASEH2C	ZN341	-1.35485	chr11-65383755	rs58950470	TF loss
183	EPHX2	MAZ	1.354589	chr8-27328354	rs2472554	TF gain
184	DLX2, DLX1	SP1	1.354506	chr2-172968775	rs62183859	TF gain
185	SLC12A6, LPCAT4, NOP10	KLF16	1.354432	chr15-34659517	rs117799466	TF gain
186	SLC12A6, LPCAT4, NOP10	KLF15	1.353072	chr15-34659517	rs117799466	TF gain
187	PLCH2, PEX10, MORN1	SPIC	1.346787	chr1-2374758	rs4648844	TF gain
188	EPHX2	ZNF398	1.345411	chr8-27328354	rs2472554	TF gain
189	SLC12A6, LPCAT4, NOP10	SP2	1.34034	chr15-34659517	rs117799466	TF gain
190	SLC12A6, LPCAT4, NOP10	SP4	1.339904	chr15-34659517	rs117799466	TF gain
191	PLCH2, PEX10, MORN1	ZNF28	1.337427	chr1-2374758	rs4648844	TF gain
192	TBL1XR1-AS1	ZN148	-1.3248	chr3-176767108	rs9825834	TF loss
193	SLC12A6, LPCAT4, NOP10	SP2	-1.3248	chr15-34659517	rs117799466	TF loss
194	SLC12A6, LPCAT4, NOP10	TBX15	1.323596	chr15-34659517	rs117799466	TF gain
195	NAB2, STAT6, NXPH4	SP3	-1.32201	chr12-57487814	rs324017	TF loss
196	MAP3K11, PCNX3, RNASEH2C	ZN467	-1.32118	chr11-65383755	rs58950470	TF loss
197	DLX2, DLX1	Zfx	-1.32086	chr2-172974789	rs62184960	TF loss

	A	B	C	D	E	F
198	PSD2	SP2	-1.32057	chr5-139028810	rs10447226	TF loss
199	DLX2, DLX1	SP3	1.320193	chr2-172968775	rs62183859	TF gain
200	NAB2, STAT6, NXPH4	VEZF1	-1.31778	chr12-57487814	rs324017	TF loss
201	EPHX2	ZBT17	1.314125	chr8-27328354	rs2472554	TF gain
202	MAP3K11, PCNX3, RNASEH2C	MAZ	-1.3129	chr11-65383755	rs58950470	TF loss
203	MAP3K11, PCNX3, RNASEH2C	ZN467	-1.30883	chr11-65383755	rs58950470	TF loss
204	KLF10	PRDM9	1.308454	chr8-103669991	rs4734654	TF gain
205	SLC12A6, LPCAT4, NOP10	HKR1	1.307877	chr15-34659517	rs117799466	TF gain
206	DLX2, DLX1, METAP1D	ZNF341	-1.30621	chr2-172955433	rs62183854	TF loss
207	TBL1XR1-AS1	VEZF1	1.30459	chr3-176767108	rs9825834	TF gain
208	ZEB2-AS1, GTDC1	ZNF823	-1.3042	chr2-145141541	rs12991836	TF loss
209	SLC12A6, LPCAT4, NOP10	KLF15	1.30312	chr15-34659517	rs117799466	TF gain