Discovering Root Causal Genes with High Throughput Perturbations

Eric V. Strobl[1,*] **and Eric R. Gamazon**[1]

[1]Vanderbilt University Medical Center, Nashville, United States of America

ABSTRACT

Root causal gene expression levels – or *root causal genes* for short – correspond to the initial changes to gene expression that generate patient symptoms as a downstream effect. Identifying root causal genes is critical towards developing treatments that modify disease near its onset, but no existing algorithms attempt to identify root causal genes from data. RNA-sequencing (RNA-seq) data introduces challenges such as measurement error, high dimensionality and non-linearity that compromise accurate estimation of root causal effects even with state-of-the-art approaches. We therefore instead leverage Perturb-seq, or high throughput perturbations with single cell RNA-seq readout, to learn the causal order between the genes. We then transfer the causal order to bulk RNA-seq and identify root causal genes specific to a given patient for the first time using a novel statistic. Experiments demonstrate large improvements in performance. Applications to macular degeneration and multiple sclerosis also reveal root causal genes that lie on known pathogenic pathways, delineate patient subgroups and implicate a newly defined omnigenic root causal model.

Root causal gene expression levels – or *root causal genes* for short – correspond to the initial changes to *gene expression* that induce a pathogenic cascade ultimately resulting in a downstream diagnosis[1,2]. Treating root causal genes can modify disease pathogenesis in its entirety, whereas targeting other causes may only provide symptomatic relief. For example, mutations in Gaucher disease cause decreased expression of wild type beta-glucocerebrosidase, or the root causal gene.[3] We can give a patient blood transfusions to alleviate the fatigue and anemia associated with the disease, but we seek more definitive treatments like recombinant glucocerebrosidase that replaces the deficient enzyme. Enzyme replacement therapy alleviates the associated liver, bone and neurological abnormalities of Gaucher disease as a downstream effect. Identifying root causal genes is therefore critical for developing treatments that eliminate disease near its pathogenic onset.

The problem is further complicated by the existence of complex disease, where a patient may have multiple root causal genes that differ from other patients even within the same diagnostic category[4]. Complex diseases often have an overwhelming number of causes, but the root causal genes may only represent a small subset implicating a more omnigenic than polygenic model. We thus also seek to identify *patient-specific* root causal genes in order to classify patients into meaningful biological subgroups each hopefully dictated by only a small group of genes.

No existing method identifies root causal genes from data. Many algorithms focus on discovering associational[5] or even causal relations[6,7], but none pinpoint the *first* gene expression levels that ultimately generate disease. We therefore define the Root Causal Strength (RCS) score to identify root causal genes unique to each patient. We then design the Root Causal Strength using Perturbations (RCSP) algorithm that estimates RCS from bulk RNA-seq under minimal assumptions by integrating Perturb-seq, or high throughput perturbation experiments using CRISPR-based technologies coupled with single cell RNA-sequencing[8–10]. Experiments demonstrate marked improvements in performance. Finally, application of the algorithm to two complex diseases with disparate pathogeneses simultaneously recovers root causal genes, omnigenic disease models, pathogenic pathways and drug candidates delineating patient subgroups for the first time.

Results

Definitions

Differential expression analysis identifies differences in gene expression levels between groups Y[5]. A gene X_i may be differentially expressed due to multiple reasons. For example, X_i may cause Y, or a confounder C may explain the relation between X_i and Y such that $X_i \leftarrow C \rightarrow Y$. In this paper, we take expression analysis a step further by pinpointing *causal* relations from expression levels regardless of the variable type of Y (discrete or continuous). We in particular seek to discover *patient-specific root causal genes* from bulk RNA-seq data, which we carefully define below.

We represent a biological system in bulk RNA-seq as a causal graph $\mathbb{G}$ – such as in Figure 1 (a) – where p vertices $\widetilde{\boldsymbol{X}}$ represent true gene expression levels in a bulk sample and Y denotes the patient symptoms or diagnosis. The set $\widetilde{\boldsymbol{X}}$ contains thousands of genes in practice. Directed edges between the vertices in $\mathbb{G}$ refer to direct causal relations. We assume that gene expression causes patient symptoms but not vice versa so that no edge from Y is directed towards $\widetilde{\boldsymbol{X}}$. The set $\text{Pa}(\widetilde{X}_i)$ refers to the *parents* of $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$, or those variables with an edge directed into $\widetilde{X}_i$. For example, $\text{Pa}(\widetilde{X}_2) = \{\widetilde{X}_1, \widetilde{X}_3\}$ in Figure 1 (a). A *root vertex* corresponds to a vertex with no parents.

We can associate $\mathbb{G}$ with the structural equation $\widetilde{X}_i = f_i(\text{Pa}(\widetilde{X}_i), E_i)$ for each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ that links each vertex to its parents and error term E_i[11]. The error term E_i is not simply a regression residual but instead represents the conglomeration of unobserved explanatory variables that only influence $\widetilde{X}_i$, such as unobserved transcriptional regulators, certain genetic variants and specific environmental conditions. We thus also include the error terms $\boldsymbol{E}$ in the directed graph of Figure 1 (b). All root vertices are error terms and vice versa. The *root causes* of Y are the error terms that cause Y, or have a directed path into Y. We define the *root causal strength* (RCS) of $\widetilde{X}_i$ on Y as the following absolute difference (Figure 1 (c)):

$$\begin{aligned}\Phi_i &= \left|\mathbb{E}(Y|\text{Pa}(\widetilde{X}_i), E_i) - \mathbb{E}(Y|\text{Pa}(\widetilde{X}_i))\right| \\ &= \left|\mathbb{E}(Y|\text{Pa}(\widetilde{X}_i), \widetilde{X}_i) - \mathbb{E}(Y|\text{Pa}(\widetilde{X}_i))\right|.\end{aligned} \qquad (1)$$

We prove the last equality in the Methods. As a result, RCS Φ_i directly measures the contribution of the gene $\widetilde{X}_i$ on Y according to its error term E_i without recovering the error term values. The algorithm does not impose functional restrictions such as additive noise to estimate the error term values as an intermediate step. Moreover Φ_i is patient-specific because the values of $\text{Pa}(\widetilde{X}_i)$ and $\widetilde{X}_i$ may differ between patients. We have $\Phi_i = 0$ when E_i is not a cause of Y, so we say that the gene $\widetilde{X}_i$ is a *patient-specific root causal gene* if $\Phi_i > 0$.

Algorithm

We propose an algorithm called Root Causal Strength using Perturbations (RCSP) that estimates $\Phi = \{\Phi_1, \ldots, \Phi_p\}$ from genes measured in both bulk RNA-seq and Perturb-seq datasets derived from possibly independent studies but from the same tissue type. We refer the reader to the Methods for details.

Estimating Φ requires access to the true gene expression levels $\widetilde{\boldsymbol{X}}$ and the removal of batch effects representing unwanted sources of technical variation such as different sequencing platforms or protocols. We however can only obtain imperfect counts $\boldsymbol{X}$ from RNA sequencing even within each batch (Figure 1 (d)). We show in the Methods that the Poisson distribution approximates the measurement error distribution induced by the sequencing process to high accuracy[12, 13]. We leverage this fact to eliminate the need for normalization by sequencing depth using an asymptotic argument where the library size N approaches infinity. N takes on a value of at least ten million in bulk RNA-seq, but we also empirically verify that the theoretical results hold well in the Supplementary Materials. We thus eliminate the Poisson measurement error and batch effects by controlling for the batches B but not N in non-linear regression models.

We in particular show that Φ_i in Equation (1) is also equivalent to:

$$\Phi_i = \left|\mathbb{E}(Y|\text{SP}(\widetilde{X}_i), X_i, B) - \mathbb{E}(Y|\text{SP}(\widetilde{X}_i), B)\right|, \qquad (2)$$

where $\text{SP}(\widetilde{X}_i)$ refers to the *surrogate parents* of $\widetilde{X}_i$, or the variables in $\boldsymbol{X}$ associated with $\text{Pa}(\widetilde{X}_i) \subseteq \widetilde{\boldsymbol{X}}$. RCSP can identify (an appropriate superset of) the surrogate parents of each variable using perturbation data because perturbing a gene changes the marginal distributions of its downstream effects – which the algorithm detects from data under mild assumptions (Figure 1 (e)). The algorithm thus only transfers the binary presence or absence of causal relations from the single cell to bulk data – rather than the exact functional relationships – in order to remain robust against discrepancies between the two data types. RCSP finally performs the two non-linear regressions needed to estimate $\mathbb{E}(Y|\text{SP}(\widetilde{X}_i), X_i, B)$ and $\mathbb{E}(Y|\text{SP}(\widetilde{X}_i), B)$ for each Φ_i. We will compare Φ_i against Statistical Dependence (SD), a measure of correlational strength defined as $\Omega_i = |\mathbb{E}(Y|X_i, B) - \mathbb{E}(Y|B)|$ where we have removed the conditioning on $\text{SP}(\widetilde{X}_i)$.

In silico identification of root causal genes

We simulated 30 bulk RNA-seq and Perturb-seq datasets from random directed graphs summarizing causal relations between gene expression levels. We performed single gene knock-down perturbations over 2500 genes and 100 batches. We obtained 200 cell samples from each perturbation, and another 200 controls without perturbations. We therefore generated a total of $2501 \times 200 = 500,200$ single cell samples for each Perturb-seq dataset. We simulated 200 bulk RNA-seq samples. We compared RCSP against the Additive Noise Model (ANM)[14, 15], the Linear Non-Gaussian Acyclic Model (LiNGAM)[1, 14], CausalCell[7], univariate regression residuals (Uni Reg), and multivariate regression residuals (Multi Reg). The first two algorithms are state-of-the-art approaches used for error term extraction and, in theory, root causal discovery. See Methods for comprehensive descriptions of the simulation setup and comparator algorithms.

We summarize accuracy results in Figure 1 (f) using the Root Mean Squared Error (RMSE) to the ground truth Φ values. All statements about pairwise differences hold true at a Bonferonni corrected threshold of 0.05/5 according to paired two-sided t-tests, since we compared a total of five algorithms. RCSP estimated Φ most accurately by a large margin. ANM and LiNGAM are theoretically correct under their respective assumptions, but they struggle to outperform standard multivariate regression due to the presence of measurement error in RNA-seq (Supplementary Materials). Feature selection and causal discovery with CausalCell did not improve performance. Univariate regression performed the worst, since it does not consider the interactions between variables. RCSP achieved the lowest RMSE while completing in about the same amount of time as multivariate regression on average (Figure 1 (g)). We conclude that RCSP both scalably and accurately estimates Φ from a combination of bulk RNA-seq and Perturb-seq data.

We will cluster the RCS values in real data to find patient subgroups. We therefore also performed hierarchical clustering using Ward's method[16] on the values of Φ estimated by RCSP with the synthetic data. We then computed the RMSE within each cluster and averaged the RMSEs between clusters. We found that RCSP maintained low average RMSE values regardless of the number of clusters considered (Figure 1 (h)). We conclude that RCSP maintains accurate estimation of Φ across different numbers of clusters.

Oxidative stress in age-related macular degeneration

We ran RCSP on a bulk RNA-seq dataset of 513 individuals with age-related macular degeneration (AMD; GSE115828)

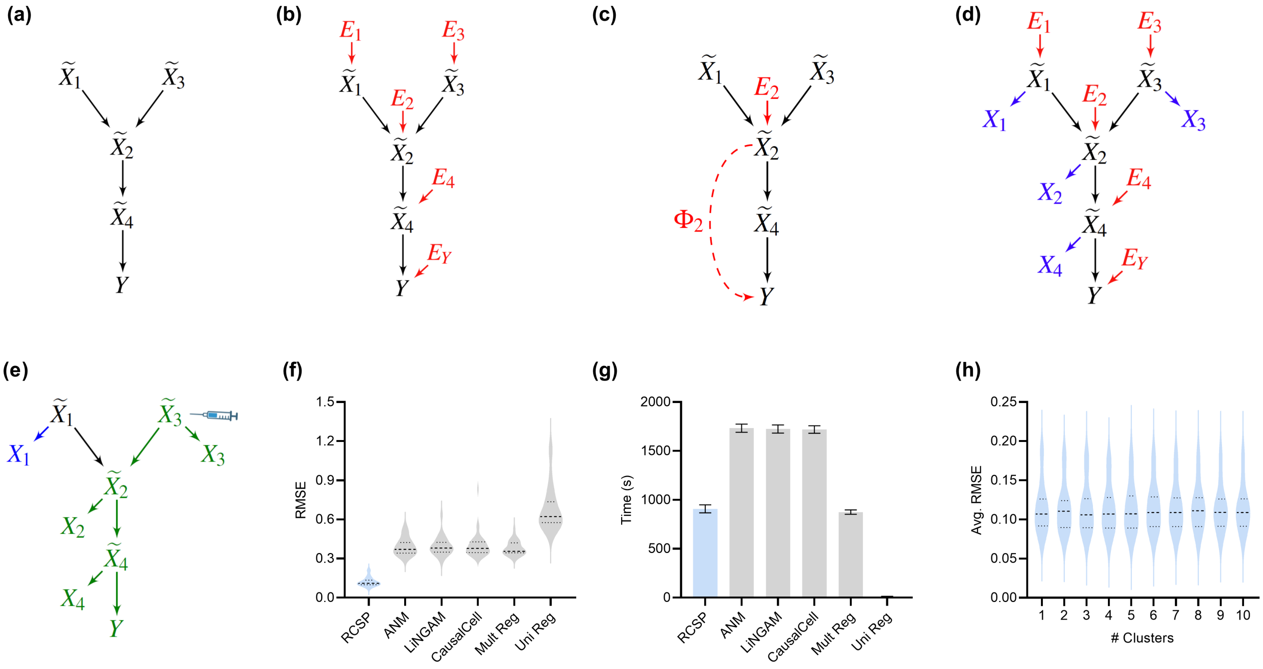


Figure 1. Method overview and synthetic data results. (a) We consider a latent causal graph over the true counts $\widetilde{\boldsymbol{X}}$. (b) We augment the graph with error terms $\boldsymbol{E}$ such that each $E_i \in \boldsymbol{E}$ in red has an edge directed towards $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$. (c) The RCS of $\widetilde{X}_2$, denoted by Φ_2, quantifies the strength of the root causal effect from E_2 to Y conditional on $\text{Pa}(\widetilde{X}_2)$. (d) We cannot observe $\widetilde{\boldsymbol{X}}$ in practice but instead observe the noisy surrogates $\boldsymbol{X}$ in blue corrupted by Poisson measurement error. (e) Perturbing a variable such as $\widetilde{X}_3$ changes the marginal distributions of downstream variables shown in green under mild conditions. (f) Violin plots show that RCSP achieved the smallest RMSE to the ground truth RCS values in the synthetic data. (g) RCSP also took about the same amount of time to complete as multivariate regression. Univariate regression only took 11 seconds on average, so its bar is not visible. Error bars denote 95% confidence intervals of the mean. (h) Finally, RCSP maintained low RMSE values regardless of the number of clusters considered.

and a Perturb-seq dataset of 247,914 cells generated from an immortalized retinal pigment epithelial (RPE) cell line[17, 18]. The Perturb-seq dataset contains knockdown experiments of 2077 genes overlapping with the genes of the bulk dataset. We set the target Y to the Minnesota Grading System score, a measure of the severity of AMD based on stereoscopic color fundus photographs. We always included age and sex as a biological variable as covariates. We do not have access to the ground truth values of Φ in real data, so we evaluated RCSP using seven alternative techniques. See Methods for a detailed rationale of the evaluation of real data. RCSP outperformed all other algorithms in this dataset (Supplementary Materials). We therefore only analyze the output of RCSP in detail here.

AMD is a neurodegenerative disease of the aging retina[19], so age is a known root cause of the disease. We therefore determined if RCSP identified age as a root cause. The algorithm estimated a heavy tailed distribution of the RCS values indicating that most of the RCS values deviated away from zero (Figure 2 (a)). The Deviation of the RCS (D-RCS), or the standard deviation from an RCS value of zero, corresponded to 0.46 – more than double that of the nearest gene (Figure 2 (d)). We conclude that RCSP correctly detected age as a root cause of AMD.

Root causal genes typically affect many downstream genes before affecting Y. We therefore expect to identify few root causal genes but many genes that correlate with Y. To evaluate this hypothesis, we examined the distribution of D-RCS relative to the distribution of the Deviation of Statistical Dependence (D-SD), or the standard deviation from an SD value of zero, in Figure 2 (b). Few D-RCS scores had large values implying the existence of only a few significant root causal genes. In contrast, most of the D-SD scores had relatively larger values concentrated around 0.10 implying the existence of many genes correlated with Y. We conclude that RCSP identified few root causal genes rather than many correlated genes for AMD.

The pathogenesis of AMD involves the loss of RPE cells. The RPE absorbs light in the back of the retina, but the combination of light and oxygen induces oxidative stress, and then a cascade of events such as immune cell activation, cellular senescence, drusen accumulation, neovascularization and ultimately fibrosis[20]. We therefore expect the root causal genes of

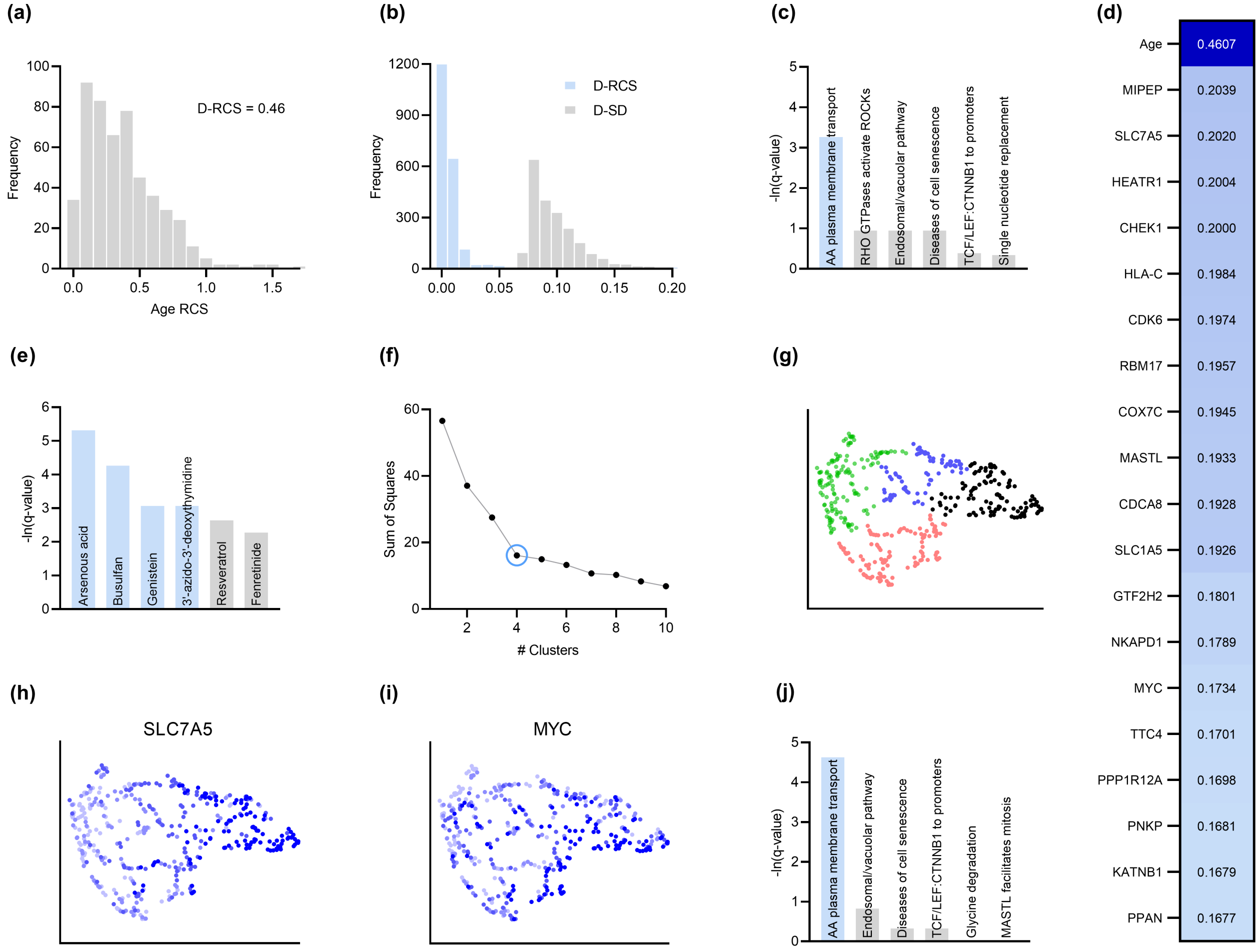


Figure 2. Analysis of AMD. (a) The distribution of the RCS scores of age deviated away from zero and had a composite D-RCS of 0.46. (b) However, the majority of gene D-RCS scores concentrated around zero, whereas the majority of gene D-SD scores concentrated around the relatively larger value of 0.10. Furthermore, the D-RCS scores of the genes in (d) mapped onto the "amino acid transport across the plasma membrane" pathway known to be involved in the pathogenesis of AMD in (c). Blue bars survived 5% FDR correction. (e) Drug enrichment analysis revealed four significant drugs, the later three of which have therapeutic potential. (f) Hierarchical clustering revealed four clear clusters according to the elbow method, which we plot by UMAP dimensionality reduction in (g). The RCS scores of the top genes in (d) increased only from the left to right on the first UMAP dimension (x-axis); we provide an example of SLC7A5 in (h) and one of three detected exceptions in (i). We therefore performed pathway enrichment analysis on the black cluster in (g) containing the largest RCS scores. (j) The amino acid transport pathway had a larger degree of enrichment in the black cluster as compared to the global analysis in (c).

AMD to include genes involved in oxidative stress. The gene MIPEP with the highest D-RCS score in Figure 2 (d) indeed promotes the maturation of oxidative phosphorylation-related proteins[21]. The second gene SLC7A5 is a solute carrier that activates mTORC1 whose hyperactivation increases oxidative stress via lipid peroxidation[22,23]. The gene HEATR1 is involved in ribosome biogenesis that is downregulated by oxidative stress[24]. The top genes discovered by RCSP thus identify pathways known to be involved in oxidative stress.

We subsequently jointly analyzed the D-RCS values of all 2077 genes. We performed pathway enrichment analysis that yielded one pathway "amino acid transport across the plasma membrane" that passed an FDR threshold of 5% (Figure 2 (c)). The leading edge genes of the pathway included the solute carriers SLC7A5 and SLC1A5. These two genes function in conjunction to increase the efflux of essential amino acids out of the lysosome[25,26]. Some of these essential amino acids like L-leucine and L-arginine activate mTORC1 that in turn increases lipid peroxidation induced oxidative stress and the subsequent degeneration of the RPE[22,23]. We conclude that pathway enrichment analysis correctly identified solute carrier genes involved in a known pathway promoting oxidative stress in AMD.

We next ran drug enrichment analysis with the D-RCS scores. The top compound arsenous acid inhibits RPE proliferation[27], but the other three significant drugs have therapeutic potential (Figure 2 (e)). Busulfan decreases the requirement for intravitreal anti-VEGF injections[28]. Genistein is a protein kinase inhibitor that similarly attenuates neovascularization[29] and blunts the effect of ischemia on the retina[30]. Finally, a metabolite of the antiviral agent 3'-azido-3'-deoxythymidine inhibits neovascularization and mitigates RPE degeneration[31]. We conclude that the D-RCS scores identified promising drugs for the treatment of AMD.

Hierarchical clustering and UMAP dimensionality reduction on the patient-specific RCS values revealed four clear clusters of patients (Figures 2 (f) and (g), respectively). Most of the top genes exhibited a clear gradation increasing only from the left to the right hand side of the UMAP embedding; we plot an example in Figure 2 (h). We found three exceptions to this rule among the top 30 genes (example in Figure 2 (i) and see Supplementary Materials). RCSP thus detected genes with large RCS scores primarily in the black cluster of Figure 2 (g). Pathway enrichment analysis within this cluster alone yielded supra-significant results on the same pathway detected in the global analysis (Figure 2 (j) versus Figure 2 (c)). Furthermore, cluster-wise drug enrichment analysis results confirmed that patients in the black cluster with many root causal genes are likely the hardest to treat (Supplementary Materials). We conclude that RCSP detected a subgroup of patients whose root causal genes have large RCS scores and involve known pathogenic pathways related to oxidative stress.

T cell infiltration in multiple sclerosis

We next ran RCSP on 137 samples collected from CD4+ T cells of multiple sclerosis (MS; GSE137143) as well as Perturb-seq data of 1,989,578 lymphoblasts, or the precursors of T cells and other lymphocytes[18,32]. We set the target Y to the Expanded Disability Status Scale score, a measure of MS severity. RCSP outperformed all other algorithms in this dataset as well (Supplementary Materials).

MS progresses over time, and RCSP correctly detected age as a root cause of MS severity with RCS values deviating away from zero (Figure 3 (a)). The distribution of gene D-RCS scores concentrated around zero, whereas the distribution of gene D-SD scores concentrated around a relatively larger value of 0.3 (Figure 3 (b)). RCSP thus detected an omnigenic root causal model rather than a polygenic correlational model.

MS is an inflammatory neurodegenerative disease that damages the myelin sheaths of nerve cells in the brain and spinal cord. T cells may mediate the inflammatory process by crossing a disrupted blood brain barrier and repeatedly attacking the myelin sheaths[33]. Damage induced by the T cells also perturbs cellular homeostasis and leads to the accumulation of misfolded proteins[34]. The root causal genes of MS thus likely include genes involved in T cell infiltration across the blood brain barrier.

Genes with the highest D-RCS scores included MNT, CERCAM and HERPUD2 (Figure 3 (d)). MNT is a MYC antagonist that modulates the proliferative and pro-survival signals of T cells after engagement of the T cell receptor[35]. Similarly, CERCAM is an adhesion molecule expressed at high levels in microvessels of the brain that increases leukocyte transmigration across the blood brain barrier[36]. HERPUD2 is involved in the endoplasmic-reticulum associated degradation of unfolded proteins[37]. Genes with the highest D-RCS scores thus serve key roles in known pathogenic pathways of MS.

We found multiple genes with high D-RCS scores in MS, in contrast to AMD where age dominated (Figure 3 (d) versus Figure 2 (d)). We performed pathway enrichment analysis using the D-RCS scores of all genes and discovered two significant pathways at an FDR corrected threshold of 5%: “adenomatous polyposis coli (APC) truncation mutants have impaired AXIN binding” and “EPH-ephrin signaling” (Figure 3 (c)). APC and AXIN are both members of the Wnt signaling pathway and regulate levels of beta-catenin[38]. Furthermore, inhibition of Wnt/beta-catenin causes CD4+ T cell infiltration into the central nervous system via the blood brain barrier in MS[39]. Ephrins similarly regulate T cell migration into the central nervous system[40] and are overexpressed in MS lesions[41]. The APC-AXIN and EPH-ephrin pathways are thus consistent with the known pathophysiology of central nervous system T cell infiltration in MS.

We subsequently performed hierarchical clustering of the RCS scores. The within cluster sum of squares plot in Figure 3 (e) revealed the presence of three clusters by the elbow method. We plot the three clusters in a UMAP embedding in Figure 3 (f). The clusters did not show a clear relationship with MS symptom severity (Supplementary Materials) or the levels of the top most genes of Figure 3 (d); we plot the MNT gene as an example in Figure 3 (g). However, further analyses with additional genes revealed that the distribution of many lower ranked genes governed the structure of the UMAP embedding (Supplementary Materials). The D-RCS scores of each cluster also implicated different mechanisms of T cell pathology including APC-AXIN in the green cluster, disturbed T cell homeostasis in the pink cluster and platelet enhanced T cell autoreactivity in the blue cluster (Supplementary Materials).

Global drug enrichment analysis did not yield any significant drugs even at a liberal FDR threshold of 10%. We thus ran drug enrichment analysis in each cluster of Figure 3 (f). The blue and pink clusters again did not yield significant drugs. However, the third green cluster identified the cysteine cathepsin inhibitors dipeptide-derived nitriles, phenylalinine derivatives, e-64, L-006235 and L-873724 (Figure 3 (h)); statistical significance of the first three held even after correcting for multiple comparisons with the Bonferroni adjustment of 0.05/4 on the q-values. The leading edge genes of the significant drugs included the cathepsins CTSL, CTSS and CTSB exclusively. These drug enrichment results corroborate

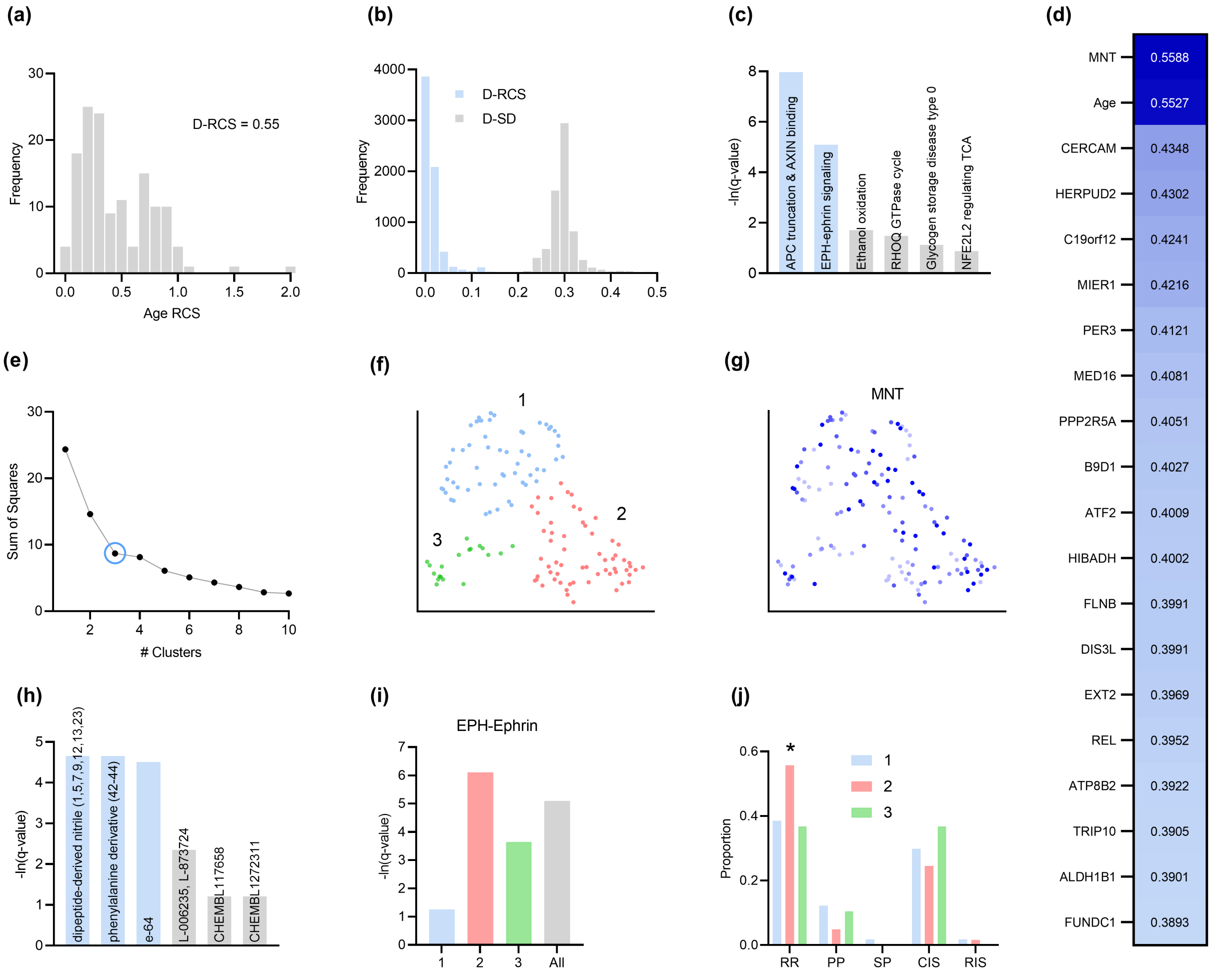


Figure 3. Analysis of MS. (a) The distribution of the RCS scores of age deviated away from zero with a composite D-RCS of 0.55. (b) The distribution of D-RCS concentrated around zero, whereas the distribution of D-SD concentrated around 0.3. (d) RCSP identified many genes with large D-RCS scores that in turn mapped onto known pathogenic pathways in MS in (c). Hierarchical clustering revealed three clusters in (e), which we plot in two dimensions with UMAP in (f). Top genes did not correlate with either dimension of the UMAP embedding; we provide an example of the MNT gene in (g). (h) Drug enrichment analysis in the green cluster implicated multiple cathepsin inhibitors. Finally, EPH-ephrin signaling survived FDR correction in (c) and was enriched in the pink cluster in (i) which contained more MS patients with the relapsing-remitting subtype in (j); subtypes include relapse-remitting (RR), primary progressive (PP), secondary progressive (SP), clinically isolated syndrome (CIS), and radiologically isolated syndrome (RIS).

multiple experimental findings highlighting the therapeutic efficacy of cathepsin inhibitors in a subgroup of MS patients responsive to interferon therapy[42,43].

Prior research has also shown that EPH-ephrin signaling is more prevalent in relapsing-remitting multiple sclerosis than in other subtypes of the disease[44]. EPH-ephrin signaling survived FDR correction in our analysis (Figure 3 (c)). Furthermore, the pathway was more enriched in the pink cluster than in the other two (Figure 3 (i)). The pink cluster indeed contained a higher proportion of patients with the relapsing-remitting subtype (Figure 3 (j)). RCSP thus precisely identified the enrichment of EPH-ephrin signaling in the correct subtype of MS.

Discussion

We presented a framework for identifying root causal genes using the error terms of structural equation models. Each error term represents the conglomeration of unobserved root causes, such as genetic variants or environmental conditions, that only modulate a specific gene. We however do not have access to many of the error terms in practice, so we introduced the root causal strength (RCS) score that detects root causal genes from bulk RNA-seq without recovering the error term values as an intermediate step. The RCSP algorithm computes RCS given knowledge of the causal ancestors of each variable, which we obtained by Perturb-seq. RCSP only transfers the causal structure (binary cause-effect relations) from the single cell to bulk data rather than the exact functional relationships in order to remain robust against discrepancies between the two data types. Results with the synthetic data demonstrated marked improvements over existing alternatives. The algorithm also recovered root causal genes that play key roles in known pathogenic pathways and implicate therapeutic drugs in both AMD and MS.

We detected a modest number of root causal genes in both AMD and MS relative to the number of genes correlated with Y. This omnigenic model differs from the omnigenic model involving *core genes*[45]. Boyle et al. define core genes as genes that directly affect disease risk and play specific roles in disease etiology. In contrast, root causal genes may not directly affect Y but lie substantially upstream of Y in the causal graph. Boyle et al. further elaborate that many *peripheral genes* affect the functions of a modest number of core genes, so the peripheral genes often explain most of disease heritability. Causation from root causal genes moves in the opposite direction – the error terms of upstream root causal genes causally affect many downstream genes that include both ancestors and non-ancestors of Y. These downstream genes contain traces of the root causal gene error terms that induce the many correlations with Y. The error terms of root causal genes associated with large RCS scores also mix with the error terms of the other ancestors of Y with small RCS scores leading to Fisher's classic *infinitesimal model*[46]. The indirect effects of root causal genes on Y and the impact of root causal genes on many downstream genes correlated with Y motivate us to use the phrase *omnigenic root causal model* in order to distinguish it from the omnigenic core gene model.

We identified root causal genes without imposing parametric assumptions using the RCS metric. Prior measures of root causal effect require restrictive functional relations, such as linear relations or additive noise, and continuous random variables[1,2,15]. These restrictions ensure exact identifiability of the underlying causal graph and error terms. However, real RNA-seq is obtained from a noisy sequencing process and contains count data arguably corrupted by Poisson measurement error[13]. The Poisson measurement error introduces confounding that precludes exact recovery of the underlying error terms. The one existing root causal discovery method that can handle Poisson measurement error uses single cell RNA-seq, estimates negative binomial distribution parameters and cannot scale to the thousands of genes required for meaningful root causal detection[47]. RCSP rectifies the deficiencies of these past approaches by ensuring accurate root causal detection even in the presence of the counts, measurement error and high dimensionality of RNA-seq.

The RCS score importantly quantifies root causal strength rather than root causal effect. As a result, the method cannot be used to identify the direction of root causal effect unconditional on the parents. The root causal effect and RCS do not differ by much in practice (Supplementary Materials), but future work should focus on exactly identifying both the strength and direction of the causal effects of the error terms.

In conclusion, RCSP integrates bulk RNA-seq and Perturb-seq to identify patient-specific root causal genes under a principled causal inference framework using the RCS score. RCS quantifies root causal strength implicitly without requiring normalization by sequencing depth or direct access to the error terms of a structural equation model. The algorithm identifies the necessary causal relations to compute RCS using reliable high throughput perturbation data rather than observational data alone. The RCS scores often suggest an omnigenic rather than a polygenic root causal model of disease. Enrichment analyses with the RCS scores frequently reveal pathogenic pathways and drug candidates. We conclude that RCSP is a novel, accurate, scalable and disease-agnostic procedure for performing patient-specific root causal discovery.

References

1. Strobl, E. V. & Lasko, T. A. Identifying patient-specific root causes of disease. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–10 (2022).

2. Strobl, E. V., Lasko, T. A. & Gamazon, E. R. Mitigating pathogenesis for target discovery and disease subtyping. *medRxiv* 2023–08 (2023).

3. Nagral, A. Gaucher disease. *J. Clin. Exp. Hepatol.* **4**, 37–50 (2014).

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. Costa-Silva, J., Domingues, D. & Lopes, F. M. Rna-seq differential expression analysis: An extended review and a software tool. *PloS One* **12**, e0190152 (2017).

6. Wang, L. *et al.* Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat. Methods* **20**, 1368–1378 (2023).

7. Wen, Y. *et al.* Applying causal discovery to single-cell analyses using causalcell. *Elife* **12**, e81464 (2023).

8. Dixit, A. *et al.* Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).

9. Adamson, B. *et al.* A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).

10. Datlinger, P. *et al.* Pooled crispr screening with single-cell transcriptome readout. *Nat. methods* **14**, 297–301 (2017).

11. Pearl, J. *Causality* (Cambridge University Press, 2009).

12. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scrna-seq. *Genome Biol.* **23**, 27 (2022).

13. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).

14. Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* (2014).

15. Strobl, E. V. & Lasko, T. A. Identifying patient-specific root causes with the heteroscedastic noise model. *J. Comput. Sci.* **72**, 102099 (2023).

16. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 236–244 (1963).

17. Ratnapriya, R. *et al.* Retinal transcriptome and eqtl analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* **51**, 606–610 (2019).

18. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell* **185**, 2559–2575 (2022).

19. Hadziahmetovic, M. & Malek, G. Age-related macular degeneration revisited: From pathology and cellular stress to potential therapies. *Front. Cell Dev. Biol.* **8**, 612812 (2021).

20. Barouch, F. C. & Miller, J. W. The role of inflammation and infection in age-related macular degeneration. *Int. ophthalmology clinics* **47**, 185–197 (2007).

21. Shi, Y. *et al.* Genetic variants at 13q12. 12 are associated with high myopia in the han chinese population. *The Am. J. Hum. Genet.* **88**, 805–813 (2011).

22. Nachef, M., Ali, A. K., Almutairi, S. M. & Lee, S.-H. Targeting slc1a5 and slc3a2/slc7a5 as a potential strategy to strengthen anti-tumor immunity in the tumor microenvironment. *Front. immunology* **12**, 624324 (2021).

23. Go, Y.-M. *et al.* Mtor-initiated metabolic switch and degeneration in the retinal pigment epithelium. *FASEB J.* **34**, 12502 (2020).

24. Turi, Z., Senkyrikova, M., Mistrik, M., Bartek, J. & Moudry, P. Perturbation of rna polymerase i transcription machinery by ablation of heatr1 triggers the rpl5/rpl11-mdm2-p53 ribosome biogenesis stress checkpoint pathway in human cells. *Cell Cycle* **17**, 92–101 (2018).

25. Nicklin, P. *et al.* Bidirectional transport of amino acids regulates mtor and autophagy. *Cell* **136**, 521–534 (2009).

26. Beaumatin, F. *et al.* mtorc1 activation requires dram-1 by facilitating lysosomal amino acid efflux. *Mol. Cell* **76**, 163–176 (2019).

27. Su, Y., Wang, F., Hu, Q., Qu, Y. & Han, Y. Arsenic trioxide inhibits proliferation of retinal pigment epithelium by downregulating expression of extracellular matrix and p27. *Int. J. Clin. Exp. Pathol.* **13**, 172 (2020).

28. Dalvin, L. A. *et al.* Busulfan treatment for myeloproliferative disease may reduce injection burden in vascular endothelial growth factor-driven retinopathy. *Am. J. Ophthalmol. Case Reports* **26**, 101554 (2022).

29. Kinoshita, S. *et al.* Genistein attenuates choroidal neovascularization. *The J. Nutr. Biochem.* **25**, 1177–1182 (2014).

30. Kamalden, T., Ji, D., Fawcett, R. & Osborne, N. Genistein blunts the negative effect of ischaemia to the retina caused by an elevation of intraocular pressure. *Ophthalmic Res.* **45**, 65–72 (2011).

31. Narendran, S. *et al.* A clinical metabolite of azidothymidine inhibits experimental choroidal neovascularization and retinal pigmented epithelium degeneration. *Investig. ophthalmology & visual science* **61**, 4–4 (2020).

32. Kim, K. *et al.* Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis. *Brain* **144**, 450–461 (2021).

33. Fletcher, J. M., Lalor, S., Sweeney, C., Tubridy, N. & Mills, K. T cells in multiple sclerosis and experimental autoimmune encephalomyelitis. *Clin. & Exp. Immunol.* **162**, 1–11 (2010).

34. Andhavarapu, S., Mubariz, F., Arvas, M., Bever Jr, C. & Makar, T. K. Interplay between er stress and autophagy: a possible mechanism in multiple sclerosis pathology. *Exp. Mol. Pathol.* **108**, 183–190 (2019).

35. Gnanaprakasam, J. R. & Wang, R. Myc in regulating immunity: metabolism and beyond. *Genes* **8**, 88 (2017).

36. Starzyk, R. M. *et al.* Cerebral cell adhesion molecule: a novel leukocyte adhesion determinant on blood-brain barrier capillary endothelium. *The J. Infect. Dis.* **181**, 181–187 (2000).

37. Kokame, K., Agarwala, K. L., Kato, H. & Miyata, T. Herp, a new ubiquitin-like membrane protein induced by endoplasmic reticulum stress. *J. Biol. Chem.* **275**, 32846–32853 (2000).

38. Spink, K. E., Polakis, P. & Weis, W. I. Structural basis of the axin–adenomatous polyposis coli interaction. *The EMBO journal* **19**, 2270–2279 (2000).

39. Lengfeld, J. E. *et al.* Endothelial wnt/β-catenin signaling reduces immune cell infiltration in multiple sclerosis. *Proc. Natl. Acad. Sci.* **114**, E1168–E1177 (2017).

40. Luo, H. *et al.* Ephrinb1 and ephrinb2 regulate t cell chemotaxis and migration in experimental autoimmune encephalomyelitis and multiple sclerosis. *Neurobiol. Dis.* **91**, 292–306 (2016).

41. Sobel, R. A. Ephrin a receptors and ligands in lesions and normal-appearing white matter in multiple sclerosis. *Brain Pathol.* **15**, 35–45 (2005).

42. Haves-Zburof, D. *et al.* Cathepsins and their endogenous inhibitors cystatins: expression and modulation in multiple sclerosis. *J. Cell. Mol. Medicine* **15**, 2421–2429 (2011).

43. Burster, T. *et al.* Interferon-γ regulates cathepsin g activity in microglia-derived lysosomes and controls the proteolytic processing of myelin basic protein in vitro. *Immunology* **121**, 82–93 (2007).

44. Golan, M. *et al.* Increased expression of ephrins on immune cells of patients with relapsing remitting multiple sclerosis affects oligodendrocyte differentiation. *Int. J. Mol. Sci.* **22**, 2182 (2021).

45. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

46. Fisher, R. A. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth Environ. Sci. Transactions Royal Soc. Edinb.* **52**, 399–433 (1919).

47. Strobl, E. V. & Lasko, T. A. Root causal inference from single cell rna sequencing with the negative binomial. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '23 (Association for Computing Machinery, New York, NY, USA, 2023).

Online Methods

Background on Causal Discovery

We denote a singleton variable like $\widetilde{X}_i$ with italics and sets of variables like $\widetilde{\boldsymbol{X}}$ with bold italics. We can represent a causal process using a *structural equation model* (SEM) linking the $p+1$ variables in $\boldsymbol{Z} = \widetilde{\boldsymbol{X}} \cup Y$ using a series of deterministic functions:

$$Z_i = f_i(\text{Pa}(Z_i), E_i), \qquad \forall Z_i \in \boldsymbol{Z} \tag{3}$$

where f_i is a function of the *parents*, or direct causes, of Z_i and an error term $E_i \in \boldsymbol{E}$. The error terms $\boldsymbol{E}$ are mutually independent. We will use the terms *vertex* and *variable* interchangeably. A *root vertex* corresponds to a vertex without any parents. On the other hand, a *terminal vertex* is not a parent of any other vertex.

We can associate a directed graph to $\boldsymbol{Z}$ by drawing a directed edge from each member of $\text{Pa}(Z_i)$ to Z_i for all $Z_i \in \boldsymbol{Z}$. A *directed path* from Z_i to Z_j corresponds to a sequence of adjacent directed edges from Z_i to Z_j. If such a path exists (or $Z_i = Z_j$), then Z_i is an *ancestor* of Z_j and Z_j is a *descendant* of Z_i. We collate all ancestors of Z_i into the set $\text{Anc}(Z_i)$. A *cycle* occurs when there exists a directed path from Z_i to Z_j and the directed edge $Z_j \rightarrow Z_i$. A *directed acyclic graph* (DAG) contains no cycles. We *augment* a directed graph by including additional vertices $\boldsymbol{E}$ and drawing a directed edge from each $E_i \in \boldsymbol{E}$ to X_i except when $X_i = E_i$ is already a root vertex. We consider an augmented DAG $\mathbb{G}$ throughout the remainder of this manuscript.

The vertices Z_i and Z_j are *d-connected* given $\boldsymbol{W} \subseteq \boldsymbol{Z} \setminus \{Z_i, Z_j\}$ in $\mathbb{G}$ if there exists a path between Z_i and Z_j such that every collider on the path is an ancestor of $\boldsymbol{W}$ and no non-collider is in $\boldsymbol{W}$. The vertices are *d-separated* if they are not d-connected. Any DAG associated with the SEM in Equation (3) also obeys the *global Markov property* where Z_i and Z_j are conditionally independent given $\boldsymbol{W}$ if they are d-separated given $\boldsymbol{W}$. The term *d-separation faithfulness* refers to the converse of the global Markov property where conditional independence implies d-separation. A distribution obeys *unconditional d-separation faithfulness* when we can only guarantee d-separation faithfulness when $\boldsymbol{W} = \emptyset$.

Causal Modeling of RNA Sequencing

Performing causal discovery requires careful consideration of the underlying generative process. We therefore propose a causal model for RNA-seq. We differentiate between the biology and the RNA sequencing technology.

We represent a biological causal process using an SEM over $\widetilde{\boldsymbol{X}} \cup Y$ obeying Equation (3). We assume that the phenotypic target Y is a terminal vertex so that gene expression causes phenotype but not vice versa. Each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ corresponds to the total number of RNA molecules of a unique gene in a single cell or bulk tissue sample. We unfortunately cannot observe $\widetilde{\boldsymbol{X}}$ in practice but instead measure a corrupted count $\boldsymbol{X}$ using single cell or bulk RNA-seq technology.

We derive the measurement error distribution from first principles. We map an exceedingly small fraction of each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ within a sample at unequal coverage. Let π_{ij} denote the probability of mapping one molecule of $\widetilde{X}_i$ in batch j so that $\sum_{i=1}^{p} \pi_{ij}$ is near zero. The law of rare events[1] implies that the Poisson distribution well-approximates the library size N so that $N \sim \text{Pois}(\sum_{i=1}^{p} \widetilde{X}_i \pi_{ij})$.

We write the probability of mapping $\widetilde{X}_i$ in a given sample as:

$$P_{ij} = \frac{\widetilde{X}_i \pi_{ij}}{\sum_{i=1}^{p} \widetilde{X}_i \pi_{ij}}.$$

This proportion remains virtually unchanged when sampling without replacement because $N \ll \sum_{i=1}^{p} \widetilde{X}_i$ with small $\sum_{i=1}^{p} \pi_{ij}$. We can therefore approximate sampling *without* replacement by sampling *with* replacement using a multinomial: $\boldsymbol{X} \sim \text{MN}(N; P_{1j}, \ldots, P_{pj})$. This multinomial and the Poisson distribution over N together imply that the marginal distribution of each $X_i \in \boldsymbol{X}$ follows an independent Poisson distribution centered at $(\sum_{i=1}^{p} \widetilde{X}_i \pi_{ij}) P_{ij} = \widetilde{X}_i \pi_{ij}$, or:

$$X_i \sim \text{Pois}(\widetilde{X}_i \pi_{ij}). \tag{4}$$

We conclude that the measurement error distribution follows a Poisson distribution to high accuracy. Multiple experimental results already corroborate this theoretical conclusion[2,12,13].

We can represent the biology and the RNA sequencing in a single DAG over $\boldsymbol{X} \cup \widetilde{\boldsymbol{X}} \cup B \cup Y$, where B denotes the batch, and Y the target variable representing patient symptoms or diagnosis. We provide a toy example in Figure 4. We draw $\mathbb{G}$ over $\boldsymbol{Z}$ in black and make each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ a parent of $X_i \in \boldsymbol{X}$ in blue. We then include the root vertex B as a parent of all members of $\boldsymbol{X}$ in green. We augment this graph with the error terms of $\widetilde{\boldsymbol{X}}$ in red and henceforth refer to the augmented DAG as $\mathbb{G}$. Repeated draws from the represented causal process generates a dataset.

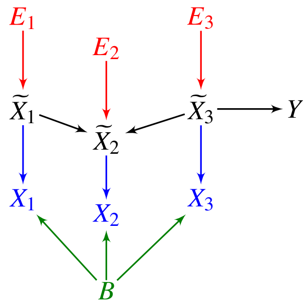


Figure 4. An example of a DAG over $\boldsymbol{X} \cup \widetilde{\boldsymbol{X}} \cup B \cup Y$ augmented with the error terms $\boldsymbol{E}$. The observed vertices $\boldsymbol{X}$ denote counts corrupted by batch B effects and Poisson measurement error.

No Need for Normalization by Sequencing Depth

We provide an asymptotic argument that eliminates the need for normalization by sequencing depth when estimating conditional expectations using bulk RNA-seq. The argument applies to the conditional expectations as a whole rather than their individual parameters.

We want to recover the causal relations between $\widetilde{\boldsymbol{X}}$ by removing batch B and depth N effects from the dataset because they correspond to the sequencing process rather than the underlying biology. We first consider removing sequencing depth by finding stably expressed housekeeping genes. Let $\widetilde{\boldsymbol{A}}$ denote the set of housekeeping genes where $\widetilde{X}_i = \widetilde{x}_i$ is a constant for each $\widetilde{X}_i \in \widetilde{\boldsymbol{A}}$; similarly $\boldsymbol{A}$ refers to the corresponding set with Poisson measurement error. Let $N = n$ be large enough such that $\sum_{X_i \in \boldsymbol{A}} x_i > 0$ for each sample. Then dividing by $L \triangleq \sum_{X_i \in \boldsymbol{A}} X_i$ controls for sequencing depth in the following sense:

$$\lim_{N\to\infty} \frac{X_i}{\sum_{X_i \in \boldsymbol{A}} X_i} = \lim_{N\to\infty} \frac{X_i/N}{\sum_{X_i \in \boldsymbol{A}} X_i/N} = \frac{P_{ij}}{\sum_{X_i \in \boldsymbol{A}} P_{ij}}$$
$$= \frac{\widetilde{X}_i \pi_{ij} / \sum_{i=1}^{p} \widetilde{X}_i \pi_{ij}}{\sum_{\widetilde{X}_i \in \widetilde{\boldsymbol{A}}} \widetilde{x}_i \pi_{ij} / \sum_{i=1}^{p} \widetilde{X}_i \pi_{ij}} = \frac{\widetilde{X}_i \pi_{ij}}{\sum_{\widetilde{X}_i \in \widetilde{\boldsymbol{A}}} \widetilde{x}_i \pi_{ij}},$$

where we have divided $\widetilde{X}_i \pi_{ij}$ by a constant in the last term. Thus, dividing by L removes measurement error within each batch as $N \longrightarrow \infty$. We assume that N is so large that the approximation error is negligible. We only invoke the assumption in bulk RNA-seq, where the library size N is on the order of at least tens of millions.

We do not divide by L in practice because we may have $L = 0$ with finite N. We instead always include $L \cup B$ in the predictor set of downstream regressions. Conditioning on $L \cup B$ ensures that all downstream regressions mitigate depth and batch effects with adequate sequencing depth, or that $\mathbb{E}(Y|\widetilde{\boldsymbol{U}}, B) = \mathbb{E}(Y|\boldsymbol{U}, L, B)$ for any $\widetilde{\boldsymbol{U}} \subseteq \widetilde{\boldsymbol{X}}$ as $N \longrightarrow \infty$. The equality holds almost surely under a mild smoothness condition:

Lemma 1. *Assume Lipschitz continuity of the conditional expectation for all $N \geq n_0$:*

$$\mathbb{E}\left|\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) - \mathbb{E}(Y|\boldsymbol{U}, L, B)\right| \leq \mathbb{E}C_N \left|\widetilde{\boldsymbol{U}} - \frac{\boldsymbol{U}}{dL}\right|,$$

where $d = \frac{\pi_{\boldsymbol{U}B}}{\sum_{\widetilde{X}_i \in \widetilde{\boldsymbol{A}}} \widetilde{x}_i \pi_{iB}}$, $C_N \in O(1)$ is a positive constant, and we have taken an outer expectation on both sides. Then $\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) = \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, L, B)$ almost surely.

We delegate proofs to the Supplementary Materials unless proven here in the Methods. Note that $\lim_{N\to\infty} \frac{\boldsymbol{U}}{dL} = \widetilde{\boldsymbol{U}}$, so the Lipschitz assumption intuitively means that accurate estimation of $\widetilde{\boldsymbol{U}}$ implies accurate estimation of $\mathbb{E}(Y|\widetilde{\boldsymbol{U}})$. Furthermore, conditioning on the library size N instead of L can introduce spurious dependencies because N depends on all of the genes rather than just the stably expressed ones.

We now eliminate the need to condition on L. Note that L is a sum of independent Poisson distributions given B per Expression (4). This implies $Y \perp\!\!\!\perp L|(\boldsymbol{U}, B)$ for any N, so that $\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) = \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, L, B) = \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, B)$ almost surely. We have proved:

Theorem 1. *Consider the same assumption as Lemma 1. Then $\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) = \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, B)$ almost surely, where we have eliminated the conditioning on L.*

We emphasize again that these equalities hold for the conditional expectation but *not* for the regression parameters; the regression parameters do not converge in general unless we divide by L. We will only need to estimate conditional expectations in order to identify root causal genes.

Identifying Root Causal Genes

We showed how to overcome Poisson measurement error without sequencing depth normalization in the previous section. We leverage this technique to define a measure for identifying the root causal genes of Y.

Definitions

A *root cause* of Y corresponds to a root vertex that is an ancestor of Y in $\mathbb{G}$. All root vertices are error terms in an augmented graph. We define the *root causal effect* of any $E_i \in \boldsymbol{E}$ on Y as $\Upsilon_i \triangleq \mathbb{P}(Y|E_i) - \mathbb{P}(Y)$[5,6].

We can identify root causes using the following result:

Proposition 1. *If $E_i \not\perp\!\!\!\perp Y$ or $E_i \not\perp\!\!\!\perp Y|\text{Pa}(\widetilde{X}_i)$ (or both), then E_i is a root cause of Y.*

We can also claim the backward direction under d-separation faithfulness. We however avoid making this additional assumption because real biological data may not arise from distributions obeying d-separation faithfulness in practice[7].

Proposition 1 implies that E_i is a root cause of Y when:

$$\Delta_i \triangleq \mathbb{P}(Y|\text{Pa}(\widetilde{X}_i), E_i) - \mathbb{P}(Y|\text{Pa}(\widetilde{X}_i)) \neq 0.$$

However, Δ_i does not correspond to the root causal effect Υ_i due to the extra conditioning on $\text{Pa}(\widetilde{X}_i)$. The two terms may also differ in direction; if $\Delta_i > 0$, then this does not imply that $\Upsilon_i > 0$, and similarly for negative values. The two variables thus represent different quantities but – in terms of priority – we would estimate Υ_i when we have nonzero Δ_i. Experimental results indicate that Υ_i and Δ_i take on similar values and agree in direction about 95% of the time in practice (Supplementary Materials).

We now encounter two challenges. First, the quantities Υ_i and Δ_i depend on the unknown error term E_i. We can however substitute E_i with $\widetilde{X}_i$ in Δ_i due to the following result:

Proposition 2. *We have* $\mathbb{P}(Y|E_i, \text{Pa}(\widetilde{X}_i)) = \mathbb{P}(Y|\widetilde{X}_i, \text{Pa}(\widetilde{X}_i))$ *under Equation* (3).

We can thus compute Δ_i without access to the error terms:

$$\begin{aligned}\Delta_i &= \mathbb{P}(Y|\text{Pa}(\widetilde{X}_i), E_i) - \mathbb{P}(Y|\text{Pa}(\widetilde{X}_i))\\ &= \mathbb{P}(Y|\text{Pa}(\widetilde{X}_i), \widetilde{X}_i) - \mathbb{P}(Y|\text{Pa}(\widetilde{X}_i)).\end{aligned}$$

The ability to determine the root causal status of E_i on Y when $\Delta_i \neq 0$ per Proposition 1, and the above ability to directly substitute E_i with its gene $\widetilde{X}_i$ both motivate the following definition: we say that $\widetilde{X}_i$ is a *root causal gene* of Y if $\Delta_i \neq 0$.

The second challenge involves computing the non-parametric probability distributions of Δ_i which come at a high cost. We thus define the analogous expected version by:

$$\begin{aligned}\Gamma_i &\triangleq \int y \left[p(y|\text{Pa}(\widetilde{X}_i), \widetilde{X}_i) - p(y|\text{Pa}(\widetilde{X}_i)) \right] dy\\ &= \mathbb{E}(Y|\text{Pa}(\widetilde{X}_i), \widetilde{X}_i) - \mathbb{E}(Y|\text{Pa}(\widetilde{X}_i))\\ &= \mathbb{E}(Y|\text{SP}(\widetilde{X}_i), X_i, B) - \mathbb{E}(Y|\text{SP}(\widetilde{X}_i), B),\end{aligned}$$

where $p(Y)$ denotes the density of Y. Observe that if $\Delta_i = 0$, then $\Gamma_i = 0$. The converse is not true but likely to hold in real data when a change in the probability distribution also changes its expectation. The set $\text{SP}(\widetilde{X}_i) \subseteq \boldsymbol{X}$ denotes the *surrogate parents* of $\widetilde{X}_i$ corresponding to the variables in $\boldsymbol{X}$ associated with $\text{Pa}(\widetilde{X}_i) \subseteq \widetilde{\boldsymbol{X}}$. The last equality holds almost surely as $N \longrightarrow \infty$ by Theorem 1.

We call $\Phi_i \triangleq |\Gamma_i|$ the *Root Causal Strength* (RCS) of $\widetilde{X}_i$ on Y. The RCS obtains a unique value $\Phi_i = \phi_{ij}$ for each patient j. We say that $\widetilde{X}_i$ is a root causal gene of Y for patient j if its RCS score is non-zero. We combine the RCS scores across a set of n samples using the Deviation of the RCS (D-RCS) $\sqrt{\frac{1}{n}\sum_{j=1}^{n}\phi_{ij}^2}$, or the standard deviation of RCS from zero. We may compute D-RCS for each cluster or globally across all patients depending on the context. We thus likewise say that $\widetilde{X}_i$ is a root causal gene for a cluster of patients or all patients in a sample if its corresponding D-RCS score for the cluster or the sample is non-zero, respectively.

Algorithm

We now design an algorithm called Root Causal Strength using Perturbations (RCSP) that recovers the RCS scores using Perturb-seq and bulk RNA-seq data.

Finding Surrogate Ancestors

Computing Φ_i for each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ requires access to the surrogate parents of each variable or, equivalently, the causal graph $\mathbb{G}$. However, inferring $\mathbb{G}$ using causal discovery algorithms may lead to large statistical errors in the high dimensional setting[8] and require restrictive assumptions such as d-separation faithfulness[9] or specific functional relations[14].

We instead directly utilize the interventional Perturb-seq data to recover a superset of the surrogate parents. We first leverage the global Markov property and equivalently write:

$$\Phi_i = \left| \mathbb{E}(Y|\text{SA}(\widetilde{X}_i), X_i, B) - \mathbb{E}(Y|\text{SA}(\widetilde{X}_i), B) \right|, \qquad (5)$$

where $\text{SA}(\widetilde{X}_i)$ denotes the *surrogate ancestors* of $\widetilde{X}_i$, or the variables in $\boldsymbol{X}$ associated with the ancestors of $\widetilde{X}_i$.

We discover the surrogate ancestors using unconditional independence tests. For any $X_k \in \boldsymbol{X}$, we test $X_k \perp\!\!\!\perp P_i$ by unpaired two-sided t-test, where P_i is an indicator function equal to one when we perturb X_i and zero in the control samples of Perturb-seq. P_i is thus a parent of X_i alone but not a child of B, so we do not need to condition on B. We use the two-sided t-test to assess for independence because the t-statistic averages over cells to mimic bulk RNA-seq. If we conclude that $X_k \not\perp\!\!\!\perp P_i$, then X_k must be a descendant of P_i by the global Markov property, so we include X_k into the set of surrogate descendants $\text{SD}(\widetilde{X}_i)$. Curating every $X_j \in \boldsymbol{X}$ such that $X_i \in \text{SD}(\widetilde{X}_j)$ into $\text{SA}(\widetilde{X}_i)$ yields the surrogate ancestors of $\widetilde{X}_i$ as desired.

Procedure

We now introduce an algorithm called Root Causal Strength using Perturbations (RCSP) that discovers the surrogate ancestors of each variable $\widetilde{\boldsymbol{X}}$ using Perturb-seq and then computes the RCS of each variable using bulk RNA-seq. We summarize RCSP in Algorithm 1.

RCSP takes Perturb-seq and bulk RNA-seq datasets as input. The algorithm first finds the surrogate descendants of each variable in $\widetilde{\boldsymbol{X}}$ in Line 2 in order to identify the surrogate ancestors of each variable in Line 5. Access to the surrogate ancestors and the batches B allows RCSP to compute Φ_i for each $X_i \in \boldsymbol{X}$ from the bulk RNA-seq in Line 6. The algorithm thus outputs the RCS scores Φ as desired.

Algorithm 1 Root Causal Strength using Perturbations (RCSP)

```
Input: bulk RNA-seq data with batches $B$, Perturb-seq data
Output: RCS scores $\Phi$
1: for each $X_i \in \boldsymbol{X}$ do
2:     $\text{SD}(\widetilde{X}_i) \leftarrow$ all $X_k \in \boldsymbol{X}$ s.t. $X_k \not\perp\!\!\!\perp P_i$ in Perturb-seq
3: end for
4: for each $X_i \in \boldsymbol{X}$ do
5:     $\text{SA}(\widetilde{X}_i) \leftarrow$ all $X_k \in \boldsymbol{X}$ s.t. $X_i \in \text{SD}(\widetilde{X}_k)$
6:     Compute $\Phi_i$ using Eq. (5) in bulk RNA-seq
7: end for
```

We certify RCSP as follows:

Theorem 2. *(Fisher consistency) Consider the same assumption as Lemma 1. If unconditional d-separation faithfulness holds, then RCSP recovers Φ almost surely as $N \to \infty$.*

We engineered RCSP to only require *unconditional* d-separation faithfulness because real distributions may not obey full d-separation faithfulness[7].

Synthetic Data

Simulations

We generated a linear SEM obeying Equation (3) specifically as $\widetilde{X}_i = \widetilde{\boldsymbol{X}}\beta_i + E_i$ for every $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ and similarly $Y = \widetilde{\boldsymbol{X}}\beta_Y + E_Y$. We included $p+1 = 2500$ variables in $\widetilde{\boldsymbol{X}} \cup Y$. We instantiated the coefficient matrix β by sampling from a Bernoulli$(2/(p-1))$ distribution in the upper triangular portion of the matrix. The resultant causal graph thus has an expected neighborhood size of 2. We then randomly permuted the ordering of the variables. We introduced weights into the coefficient matrix by multiplying each entry in β by a weight sampled uniformly from $[-1,-0.25] \cup [0.25,1]$. The error terms each follow a standard Gaussian distribution multiplied by 0.5. We introduced batch effects by drawing each entry of the mapping efficiencies π from the uniform distribution between 10 and 1000 for the bulk RNA-seq, and between 0.1 and 1 for the Perturb-seq. We set $\widetilde{X}_i \leftarrow \text{softplus}(\widetilde{X}_i)$ and then obtained the corrupted surrogate X_i distributed Pois$(\widetilde{X}_i\pi_{ij})$ for each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ and batch j. We chose Y uniformly at random from the set of vertices with at least one parent and no children. We repeated the above procedure 30 times.

Comparators

We compared RCSP against the following four algorithms:

(1) Additive noise model (ANM)[14,15]: performs non-linear regression of X_i on Pa$(X_i) \cup B$ and then regresses Y on the residuals $\boldsymbol{E} \setminus E_i$ to estimate $|\mathbb{E}(Y|\boldsymbol{E} \setminus E_i) - \mathbb{E}(Y|\boldsymbol{X}, B)|$ for each $X_i \in \boldsymbol{X}$. The non-linear regression residuals are equivalent to the error terms assuming an additive noise model.

(2) Linear Non-Gaussian Acyclic Model (LiNGAM)[1,14]: same as above but performs linear instead of non-linear regression.

(3) CausalCell[7]: selects the top 50 genes with maximal statistical dependence to Y, and then runs the PC algorithm using a non-parametric conditional independence test to identify a causal graph among the top 50 genes. The algorithm finally performs root causal inference with ANM as above but uses the estimated parent sets for the top 50 genes and the Perturb-seq data otherwise.

(4) Univariate regression residuals (Uni Reg): regresses Y on $X_i \cup B$ and estimates the absolute residuals $|Y - \mathbb{E}(Y|X_i, B)|$ for each $X_i \in \boldsymbol{X}$.

(5) Multivariate regression residuals (Multi Reg): similar to above but instead computes the absolute residuals after regressing Y on $(\boldsymbol{X} \setminus X_i) \cup B$.

The first two methods are state-of-the-art approaches used for root causal discovery. Univariate and multivariate regressions do not distinguish between predictivity and causality, but we included them as sanity checks. We performed all non-linear regressions using multivariate adaptive regression splines to control for the underlying regressor[14]. We compared the algorithms on their accuracy in estimating Φ.

Real Data

Quality Control

We downloaded Perturb-seq datasets of retinal pigment epithelial cells from the RPE-1 cell line, and lymphoblast cells from the K562 cell line[18]. We used the genome-wide dataset version for the latter. We downloaded the datasets from the scPerturb database on Zenodo[16] with the same quality controls as the original paper. Replogle et al. computed adjusted library sizes by equalizing the mean library size of control cells within each batch. Cells with greater than a 2000 or 3000 library size, and less than 25% or 11% mitochondrial RNA were kept, respectively. The parameters were chosen by plotting the adjusted library sizes against the mitochondrial RNA counts and then manually setting thresholds that removed low quality cells likely consisting of ambient mRNA transcripts arising from premature cell lysis or cell death.

We next downloaded bulk RNA-seq datasets derived from patients with age-related macular degeneration (AMD; GSE115828) and multiple sclerosis (MS; GSE137143)[17,32]. We excluded 10 individuals from the AMD dataset including one with an RNA integrity number of 21.92, five missing an integrity number (all others had an integrity number of less than 10), and four without a Minnesota Grading System score. We kept all samples from the MS dataset derived from CD4+ T cells but filtered out genes with a mean of less than 5 counts as done in the original paper.

We finally kept genes that were present in both the AMD bulk dataset and the RPE-1 Perturb-seq dataset, yielding a final count of 513 bulk RNA-seq samples and 247,914 Perturb-seq samples across 2077 genes. We also kept genes that were present in both the MS bulk dataset and the K562 Perturb-seq dataset, yielding a final count of 137 bulk RNA-seq samples and 1,989,578 Perturb-seq samples across 6882 genes. We included age and sex as a biological variable as covariates for every patient in both datasets in subsequent analyses.

Evaluation Rationale

We do not have access to the ground truth values of Φ in real data. We instead evaluate the RCSP estimates of Φ using alternative sources of ground truth knowledge. We first assess the accuracy of RCS using the control variable age as follows:

(1) Determine if the RCS values of age identify age as a root cause in diseases that progress over time.

Second, root causal genes should imply a more omnigenic than polygenic model because the effects of a few error terms distribute over many downstream genes. We verify omnigenecity as follows:

(2) Determine if the distribution of D-RCS concentrates around zero more than the distribution of the Deviation of Statistical Dependence (D-SD) defined as $\sqrt{\frac{1}{n}\sum_{j=1}^{n}\omega_{ij}^2}$ for each gene $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$ where $\Omega_i = |\mathbb{E}(Y|X_i, B) - \mathbb{E}(Y|B)|$ and ω_{ij} its value for patient j.

Despite the sparsity/omnigenecity of root causal genes, we still expect the root causal genes to correspond to at least some known causes of disease:

(3) Determine if genes with the top D-RCS scores correspond to genes known to cause the disease.

Next, the root causal genes initiate pathogenesis, and we often have knowledge of pathogenic pathways even though we may not know the exact gene expression cascade leading to disease. Intervening on root causal genes should also modulate patient symptoms. We thus further evaluate the accuracy of RCSP using pathway and drug enrichment analyses as follows:

(4) Determine if the D-RCS scores identify known pathogenic pathways of disease in pathway enrichment analysis.

(5) Determine if the D-RCS scores identify drugs that treat the disease.

Finally, complex diseases frequently involve multiple pathogenic pathways that differ between patients. Patients with the same complex disease also respond differently to treatment. We hence evaluate the precision of RCS as follows:

(6) Determine if the patient-specific RCS scores identify subgroups of patients involving different but still known pathogenic pathways.

(7) Determine if the patient-specific RCS scores identify subgroups of patients that respond differently to drug treatment.

In summary, we evaluate RCSP in real data based on its ability to (1) identify age as a known root cause, (2) suggest an omnigenic root causal model, (3) recover known causal genes, (4) find known pathogenic pathways, (5) find drugs that treat the disease, and (6,7) delineate patient subgroups.

Enrichment Analyses

Multivariate adaptive regression splines introduce sparsity, but enrichment analysis performs better with a dense input. We can estimate the conditional expectations of Φ using any general non-linear regression method, so we instead estimated the expectations using kernel ridge regression equipped with a radial basis function kernel[19]. We then computed the D-RCS across all patients for each variable in $\boldsymbol{X}$. We ran pathway enrichment analysis using the fast gene set enrichment analysis (FGSEA) algorithm[20] with one hundred thousand simple permutations using the D-RCS scores and pathway information from the Reactome database (version 1.86.0)[21]. We likewise performed drug set enrichment analysis with the Drug Signature database (version 1.0)[22]. We repeated the above procedures for the D-RCS of any cluster identified by hierarchical clustering via Ward's method[16].

Data Availability

All datasets analyzed in this study have been previously published and are publicly accessible as follows:

1. Bulk RNA-seq for AMD: GSE115828
2. Bulk RNA-seq for MS: GSE137143
3. Perturb-seq for the RPE-1 and K562 cell lines: DOI 10044268

Code Availability

R code needed to replicate all experimental results is available on GitHub.

Acknowledgements

Research reported in this report was supported by the National Human Genome Research Institute of the National Institutes of Health under award numbers R01HG011138 and R35HG010718.

Online Methods References

1. Papoulis, A. *Probability, Random Variables and Stochastic Processes* (McGraw-Hill, 1984).
2. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
3. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).
4. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scrna-seq. *Genome Biol.* **23**, 27 (2022).
5. Strobl, E. V. Counterfactual formulation of patient-specific root causes of disease. *J. Biomed. Informatics* (2024).
6. Strobl, E. V. & Lasko, T. A. Sample-specific root causal inference with latent variables. In *Conference on Causal Learning and Reasoning*, 895–915 (PMLR, 2023).
7. Strobl, E. V. Causal discovery with a mixture of dags. *Mach. Learn.* 1–25 (2022).
8. Colombo, D., Maathuis, M. H. *et al.* Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15**, 3741–3782 (2014).
9. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* (MIT press, 2000), 2nd edn.
10. Peters, J., Mooij, J. M., Janzing, D. & Schölkopf, B. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* (2014).
11. Strobl, E. V. & Lasko, T. A. Identifying patient-specific root causes with the heteroscedastic noise model. *J. Comput. Sci.* **72**, 102099 (2023).

12. Strobl, E. V. & Lasko, T. A. Identifying patient-specific root causes of disease. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–10 (2022).

13. Wen, Y. *et al.* Applying causal discovery to single-cell analyses using causalcell. *Elife* **12**, e81464 (2023).

14. Friedman, J. H. Multivariate adaptive regression splines. *The Annals Stat.* **19**, 1–67 (1991).

15. Replogle, J. M. *et al.* Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell* **185**, 2559–2575 (2022).

16. Green, T. D. *et al.* scperturb: Information resource for harmonized single-cell perturbation data. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life* (2022).

17. Ratnapriya, R. *et al.* Retinal transcriptome and eqtl analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* **51**, 606–610 (2019).

18. Kim, K. *et al.* Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis. *Brain* **144**, 450–461 (2021).

19. Shawe-Taylor, J. & Cristianini, N. *Kernel Methods for Pattern Analysis* (Cambridge University Press, 2004).

20. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* **60012**, 1–9 (2016).

21. Fabregat, A. *et al.* Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinforma.* **18**, 1–9 (2017).

22. Yoo, M. *et al.* Dsigdb: drug signatures database for gene set analysis. *Bioinformatics* **31**, 3069–3071 (2015).

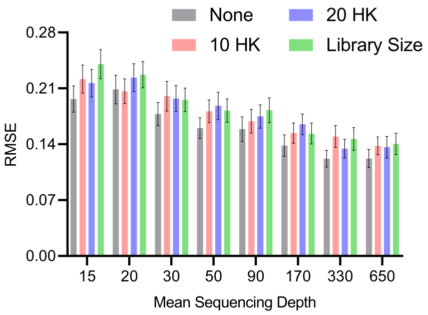
23. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 236–244 (1963).

Supplementary Materials

Normalization by Sequencing Depth

We theoretically showed that RCS does not require normalization by sequencing depth in the Methods using an asymptotic argument. We tested this claim empirically by drawing 200 bulk RNA-seq samples from random DAGs as in the Methods but over $p+1 = 250$ variables. We varied the mean sequencing depth N/p of each gene from 15, 20, 30, 50, 90, 170, 330 to 650 counts; multiplying N/p by p recovers the library size N. We only included one batch in the bulk RNA-seq in order to isolate the effect of sequencing depth. We compared no normalization, normalization by 10 housekeeping genes, normalization by 20 housekeeping genes, and normalization by library size. We repeated each experiment 100 times and thus generated a total of $100 \times 4 \times 8 = 3200$ datasets.

We plot the results in Supplementary Figure 1. All methods improved with increasing mean sequencing depth as expected. The no normalization strategy performed the best at low mean sequencing depths, followed by the housekeeping genes and then total library size. The result even held with a small library size of $N = 15 \times 249 = 3735$ at the smallest mean sequencing depth of 15, suggesting that the asymptotic argument holds well in bulk RNA-seq where N/p is often greater than 500 and N greater than the tens of millions. However, the average RMSEs of all normalization methods became more similar as sequencing depth increased. We conclude that normalization by sequencing depth exceeds or matches the accuracy of other strategies. We therefore do not normalize by sequencing depth in subsequent analyses.

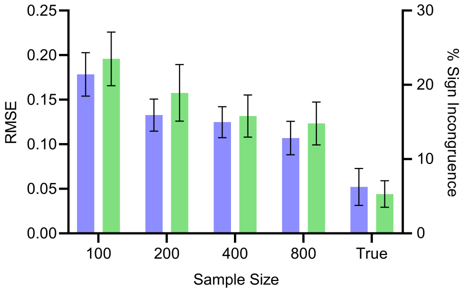


Supplementary Figure 1. Mean RMSE to the ground truth RCS values across different mean sequencing depths and normalization strategies. The no normalization strategy achieved low RMSEs at lower mean sequencing depths, but the performances of all methods converged as the mean sequencing depths increased. Error bars denote 95% confidence intervals of the mean RMSE.

Root Causal Effect versus Signed Root Causal Strength

We compared the root causal effects Γ and the signed RCS, or Δ. The two quantities are not equivalent, but they are similar. We empirically investigated the differences between the estimated values of Δ and the true values of Γ using the RMSE and also the percent of samples with incongruent signs; Δ and Γ have incongruent signs if one is positive and the other is negative. We again drew 200 bulk RNA-seq samples from random DAGs as in the Methods over $p+1 = 250$ variables with one batch. We varied the bulk RNA-seq sample size from 100, 200, 400 to 800. We also compared true Δ against true Γ by estimating the two to negligible error using 20,000 samples of $\widetilde{\boldsymbol{X}}$. We repeated each experiment 100 times and thus generated a total of $100 \times 5 = 500$ datasets.

We summarize the results in Supplementary Figure 2. The estimated Δ values approached the true Γ values with increasing sample sizes. The true Δ values did not converge exactly to the true Γ values, but the RMSE remained low at 0.05 and the two values differed in sign only around 5.3% of the time. Increasing the number of samples of $\widetilde{\boldsymbol{X}}$ to 50,000 did not change performance, confirming that we reached the floor. We conclude that the empirical results replicate the theoretical results because Δ and Γ do not match exactly. However, the two quantities take on similar values and their signs matched around 95% of the time in practice.

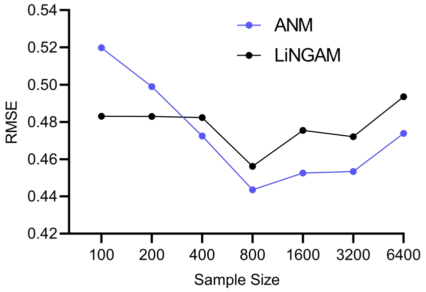


Supplementary Figure 2. Mean RMSE and percent sign incongruence of RCE and signed RCS values. The RMSE continues to decrease with increasing sample size but reaches a floor of around 0.05. Similarly, the percent sign incongruence decreases but reaches a floor of around 5%.

Functional Causal Models and Measurement Error

The experiments in the Results section quantify the accuracies of the algorithms in estimating Φ. However, the functional causal models ANM and LiNGAM also estimate the error terms as an intermediate step, whereas RCSP does not. We therefore also investigated the accuracies of ANM and LiNGAM in estimating the error term values.

Theoretical results suggest that ANM and LiNGAM cannot consistently estimate the error terms in RNA-seq due to the Poisson measurement error. We empirically tested this hypothesis by sampling from bulk RNA-seq data as in the Methods but with $p+1 = 100$ and a batch size of one in order to isolate the effect of measurement error. We repeated the experiment 100 times for bulk RNA-seq sample sizes of 100, 200, 400, 800, 1600 and 3200. We plot the results in Supplementary Figure 3. The accuracies of ANM and LiNGAM did not improve beyond an RMSE of 0.44 even with a large sample size of 6400. We conclude that ANM and LiNGAM cannot estimate the error terms accurately in the presence of measurement error even with large sample sizes.



Supplementary Figure 3. Mean RMSE values to the ground truth error term values across different sample sizes. The accuracies of ANM and LiNGAM do not improve with increasing sample sizes.

Additional Results for Age-related Macular Degeneration

Algorithm Comparisons

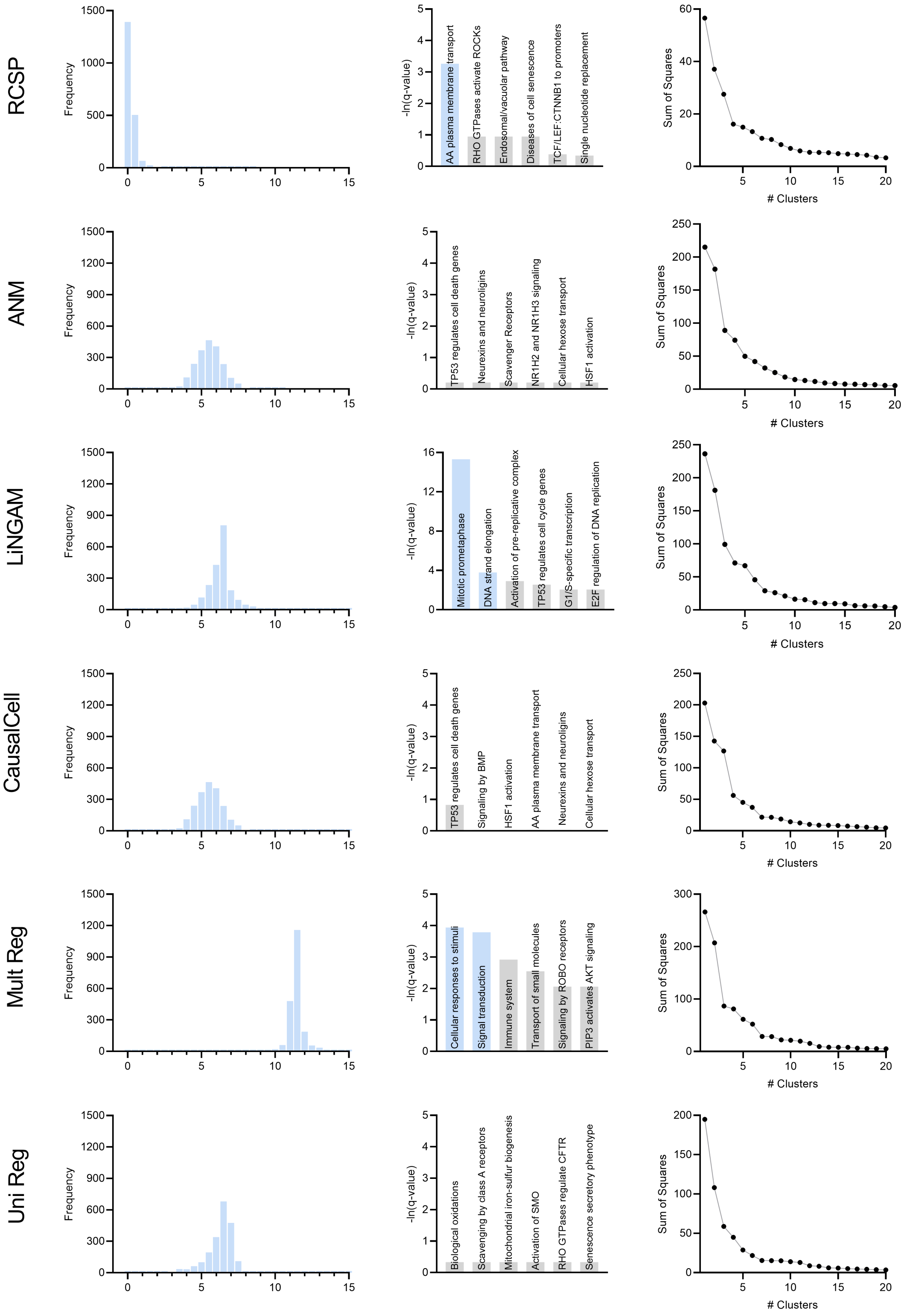
We say that an algorithm performs well in real data if it simultaneously (1) identifies an omnigenic model, (2) recovers known pathogenic pathways with high specificity measured by the sparsity of leading edge genes, and (3) clusters patients into clear subgroups.

We compared the algorithms with the AMD data. We summarize the results in Supplementary Figure 4 plotted on the next page. The first column denotes the standard deviation of the outputs for each algorithm. We standardized the outputs to have mean zero unit variance, and then added the minimum value so that all histograms begin at zero. Only the RCSP algorithm had a histogram with large probability mass centered around zero. Incorporating feature selection and causal discovery with CausalCell introduced more outliers in the histogram of ANM. We conclude that only RCSP detected an omnigenic root causal model.

We plot the results of pathway enrichment analysis in the second column of Supplementary Figure 4. RCSP, LiNGAM and univariate regression detected pathways related to oxidative stress in AMD. However, the "mitotic prometaphase" and "DNA strand elongation" pathways in blue for LiNGAM involved 94 and 27 leading edge genes, respectively. The "cellular responses to stimuli" and "signal transduction" pathways for multivariate regression also involved 253 and 282 leading edge genes. In contrast, the "amino acid plasma membrane transport" pathway for RCSP involved two leading edge genes. We conclude that RCSP identified a known pathogenic pathway of AMD with the fewest number of leading edge genes.

We finally plot the clustering results in the third column of Supplementary Figure 4. The RCSP sum of squares plot revealed four clear groups of patients, whereas the other plots did not reveal a clear number of categories using the elbow method. We conclude that only RCSP identified clear subgroups of patients in AMD.

In summary, RCSP detected the most omnigenic model, identified pathogenic pathways with maximal specificity and discovered distinguishable patient subgroups. We therefore conclude that RCSP outperformed all other algorithms in the AMD dataset.



Supplementary Figure 4. Comparison of the algorithms in age-related macular degeneration.

Biological Results

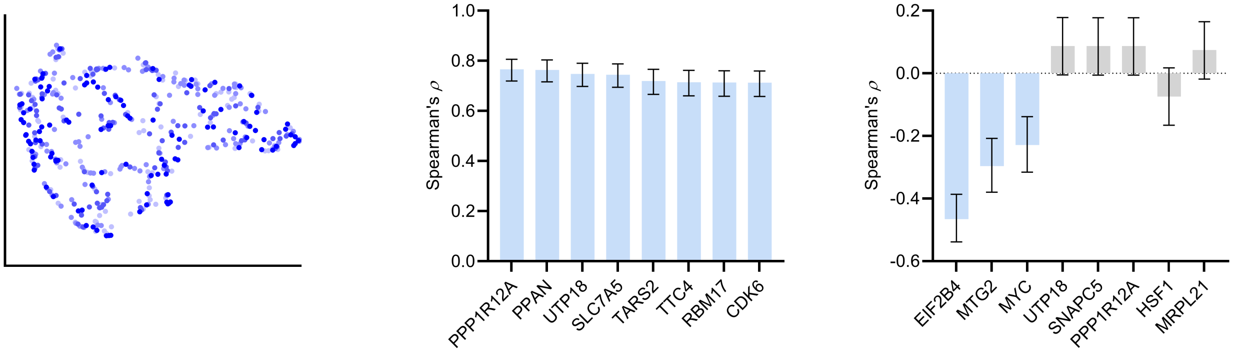
We provide the full pathway enrichment analysis results in Supplementary Table 1 corresponding to Figure 2 (c). We summarize pathway enrichment analysis of the black cluster of Figure 2 (g) in Figure 2 (j). However, analyses of the blue, green and pink clusters did not yield significant pathways even at a liberal FDR threshold of 10%.

Pathway	p-value	q-value	Effect Size	Leading Edge
Amino acid transport across the plasma membrane	2.44e-05	0.038	0.995	8140,6510
RHO GTPases Activate ROCKs	2.09e-03	0.388	0.976	4659,5500
Endosomal/Vacuolar pathway	2.32e-03	0.388	0.998	3107
Diseases of Cellular Senescence	2.97e-03	0.388	0.997	1021
Binding of TCF/LEF:CTNNB1 to target gene promoters	6.52e-03	0.680	0.993	4609
APEX1-Indep. Resolution of AP Sites via Nucleotide Replacement	7.28e-03	0.712	0.980	11284,7515
MASTL Facilitates Mitotic Progression	1.59e-02	0.978	0.911	84930,983
PI5P Regulates TP53 Acetylation	1.94e-02	0.978	0.980	79837
Formation of Incision Complex in GG-NER	2.24e-02	0.978	0.791	2966,9978,2967
Glycine degradation	2.24e-02	0.978	0.977	1738
Prefoldin mediated transfer of substrate to CCT/TriC	3.96e-02	0.978	0.787	5203,5201,10576

Supplementary Table 1. Full pathway enrichment analysis results for all patients in the AMD dataset. We list the Entrez gene IDs of up to the top three leading edge genes in the right-most column.

We examined whether the clusters of Figure 2 (g) differentiate dry and wet macular degeneration. Wet macular degeneration is associated with the highest Minnesota Grading System (MGS) score of 4[1]. We plotted the UMAP embedding against MGS (Supplementary Figure 5 (a)). None of the two UMAP dimensions correlated significantly with the MGS score (5% uncorrected threshold by Spearman's correlation test). These results and the large RCS scores of age in Figure 2 (a) seem to support the hypothesis that wet macular degeneration is a more severe type of dry macular degeneration. However, MGS does not differentiate between wet macular degeneration and late stage dry macular degeneration involving geographical atrophy. We therefore cannot separate late stage dry and wet macular degeneration using the RCS scores alone.

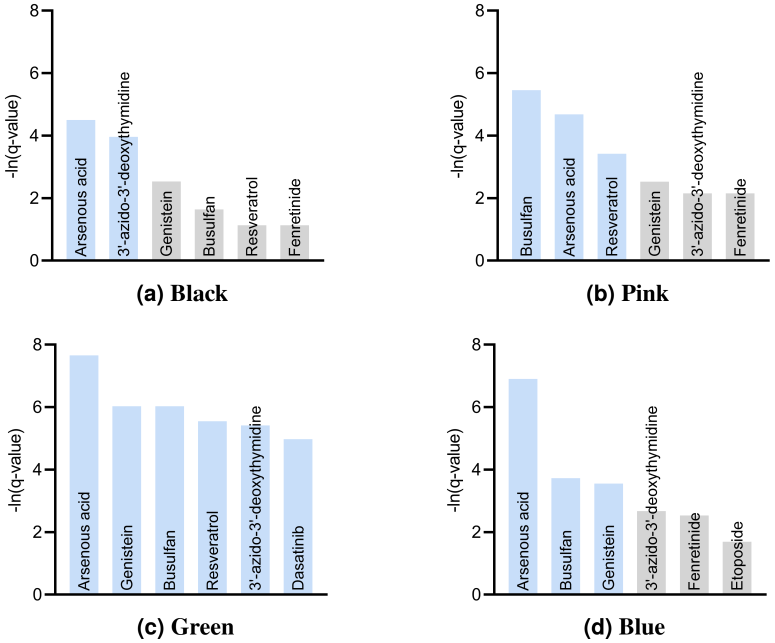
We correlated the two UMAP dimensions with the top 30 genes ranked by their RCS scores. We plot genes with the highest correlation to the first and second UMAP dimensions in Supplementary Figures 5 (b) and 5 (c), respectively. Many genes correlated with the first dimension, but only three genes correlated with the second at an FDR threshold of 5%.



(a) Severity **(b) UMAP Dimension 1** **(c) UMAP Dimension 2**

Supplementary Figure 5. Additional UMAP embedding results for AMD. (a) The UMAP dimensions did not correlate with AMD severity as assessed by the MGS score. Many genes correlated with the first UMAP dimension in (b), but only three genes correlated with the second UMAP dimension in (c). Blue bars passed an FDR threshold of 5%, and error bars denote 95% confidence intervals.

We finally performed drug enrichment analysis in each of the four clusters in Figure 2 (g). We summarize the results in Supplementary Figure 6. Only two drugs – and one potentially therapeutic option – passed FDR correction in patients in the black cluster with the most identified root causal genes according to the RCS scores. In contrast, enrichment analysis identified many drugs in patients in the green cluster with the lowest RCS scores and thus relatively few root causal genes. The pink and blue clusters yielded moderate results. We conclude that drug enrichment analysis expectedly identified more drugs for patients on the left hand side of the UMAP embedding with fewer root causal genes than on the right hand side with many simultaneous root causal genes.



Supplementary Figure 6. Drug enrichment analysis results by cluster in Figure 2 (g). The analyses recovered similar drugs across clusters, but the results for the green cluster in (c) were supra-significant.

Additional Results for Multiple Sclerosis

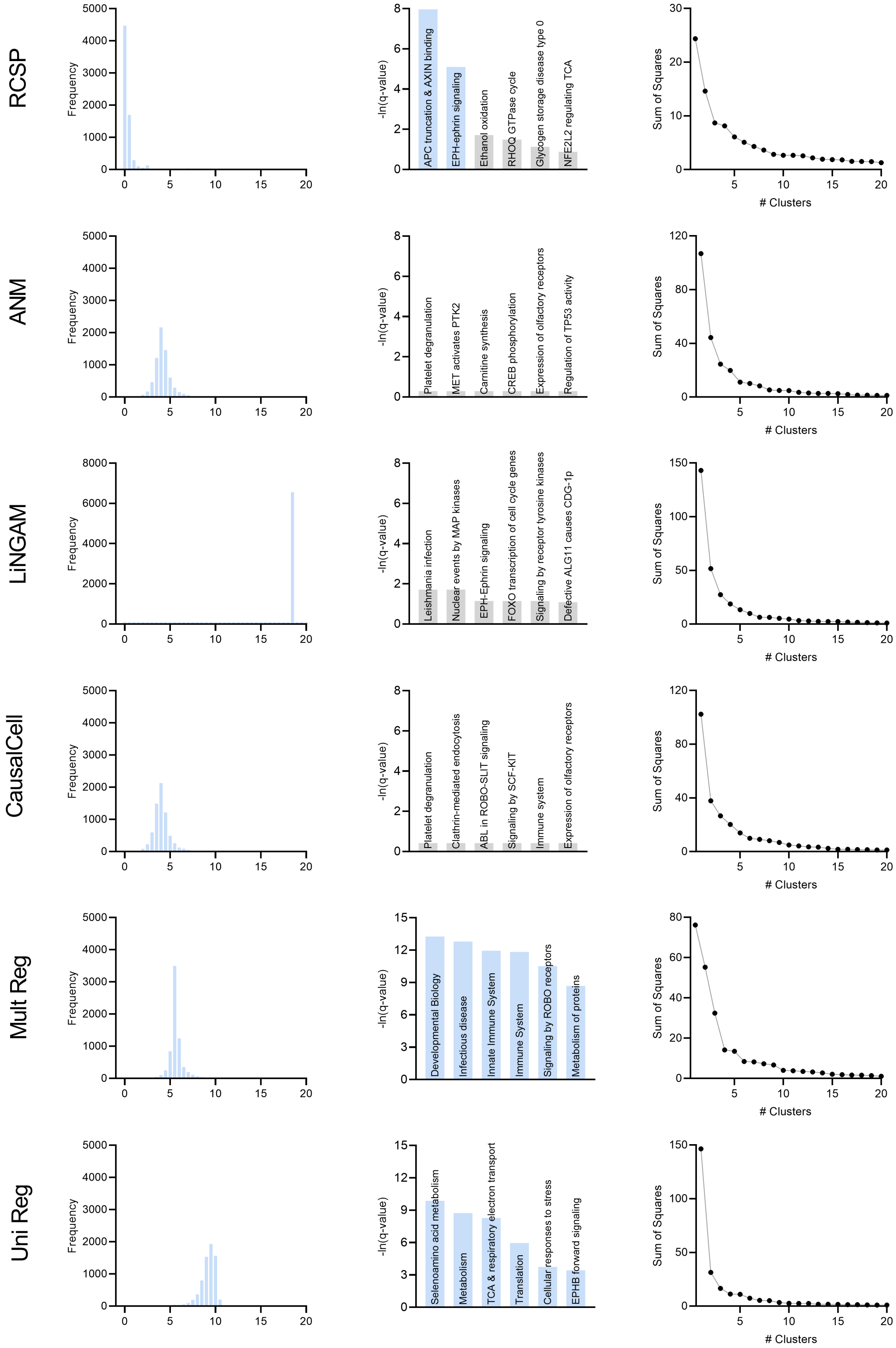
Algorithm Comparisons

We compared the algorithms using the MS data with the same criteria used for the AMD dataset. We summarize the results in Supplementary Figure 7 plotted on the next page. Only the histogram of RCSP had large probability mass centered around zero as shown in the first column. The histogram of LiNGAM contained many outliers, so it appears to spike around a value of 18. The histograms of ANM and CausalCell were again near identical. We conclude that only the histogram of RCSP supported an omnigenic root causal model in MS.

We performed pathway enrichment analysis on the algorithm outputs and summarize the results in the second column of Supplementary Figure 7. The functional causal models ANM, LiNGAM and CausalCell did not identify significant pathways at an FDR corrected threshold of 0.05. In contrast, multivariate and univariate regression both identified many significant pathways in blue with no specific link to the blood brain barrier. The top six significant pathways for multivariate and univariate regression involved 112 to 831 and 18 to 545 leading edge genes, respectively. In contrast, the two significant pathways of RCSP involved only 2 and 9 leading genes. We conclude that RCSP detected pathogenic pathways of MS with the sparsest set of leading edge genes.

We finally clustered the algorithm outputs into patient subgroups. We list the sum of squares plots in the third column of Supplementary Figure 7. Univariate regression did not differentiate between the patients because it detected one dominating cluster. RCSP and multivariate regression identified clear subgroups according to the elbow method, whereas the sum of squares plots for ANM, LiNGAM and CausalCell showed no clear cutoffs. We conclude that only RCSP and multivariate regression identified clear patient subgroups in MS.

In summary, only RCSP simultaneously detected an omnigenic root causal model, identified pathogenic pathways with high specificity and discovered clear patient subgroups. We therefore conclude that RCSP also outperformed all other algorithms in the MS dataset.



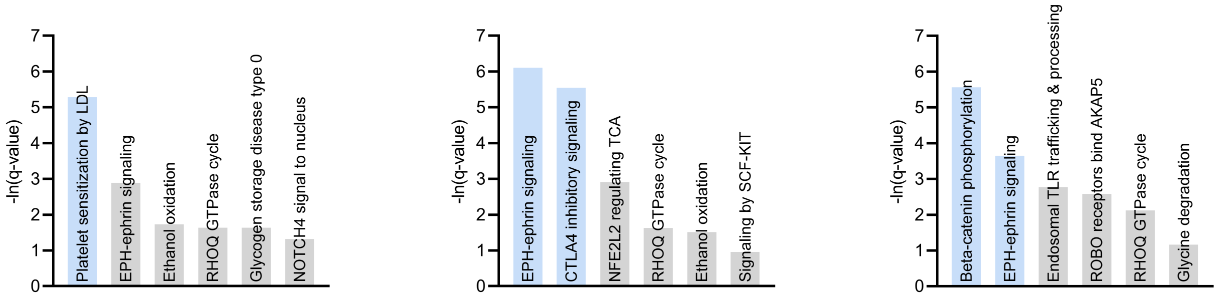
Supplementary Figure 7. Comparison of the algorithms in multiple sclerosis.

Biological Results

We provide the full global pathway enrichment analysis results for MS in Supplementary Table 2. Pathway enrichment analysis of the individual clusters in Figure 3 (f) consistently implicated EPH-ephrin signaling among the top two pathways. However, each cluster also involved one separate additional pathway. The green cluster involved the same APC-AXIN pathway as the global analysis via beta-catenin. On the other hand, the blue cluster involved "platelet sensitization by LDL." Low density lipoprotein enhances platelet aggregation. Platelet degranulation in turn drives the generation of autoreactive T cells in the peripheral circulation during disturbance of the blood brain barrier[2]. Finally, CTLA4 regulates T-cell homeostasis and inhibits autommunity for the pink cluster[3]. The D-RCS scores of each cluster thus implicate different mechanisms of T cell pathology.

Pathway	p-value	q-value	Effect Size	Leading Edge
APC truncation mutants have impaired AXIN binding	1.91e-06	3.45e-4	0.960	5525,5527
EPH-ephrin signaling	4.23e-05	6.12e-3	0.826	8874,102,8976
Ethanol oxidation	2.02e-03	0.182	0.967	219,128
RHOQ GTPase cycle	2.72e-03	0.226	0.793	9322,8874,10395
Glycogen storage disease type 0 (muscle GYS1)	4.32e-03	0.322	0.996	2992
NFE2L2 regulating TCA cycle genes	6.31e-03	0.414	0.970	4199,3417
C6 deamination of adenosine	7.42e-03	0.414	0.981	103,104
Ion channel transport	7.63e-03	0.414	0.728	57198,540,55515
Synthesis of IP3 and IP4 in the cytosol	7.65e-03	0.414	0.904	3633,805,23236
Diseases associated with glycosaminoglycan metabolism	8.21e-03	0.414	0.894	2132,11285,3339
Signaling by SCF-KIT	8.67e-03	0.414	0.794	7006,5578,3815

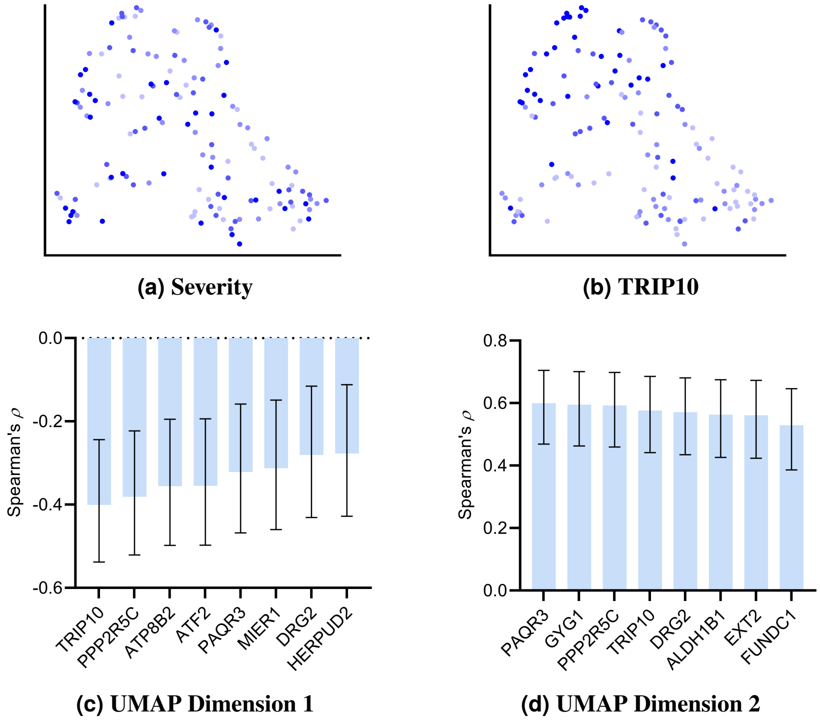
Supplementary Table 2. Full pathway enrichment analysis results for all patients in the MS dataset. We again list up to the top three leading edge genes in the right-most column.



(a) Blue **(b) Pink** **(c) Green**

Supplementary Figure 8. Pathway enrichment analysis results by cluster consistently revealed EPH-ephrin signaling as well as an additional pathway implicating T cell pathology.

The severity of MS, as assessed by the Expanded Disability Status Scale (EDSS) score, did not correlate with either dimension of the UMAP embedding (Supplementary Figure 9 (a)). The top genes in Figure 3 (d) such as MNT and CERCAM also did not correlate. However, lower ranked genes such as TRIP10 did (Supplementary Figure 9 (b)). An expanded correlation analysis with the top 30 genes revealed significant correlations across a variety of lower ranked genes (Supplementary Figures 9 (c) and 9 (d)). We conclude that the distribution of lower ranked genes govern the structure of the UMAP embedding in Figure 3 (f).



(a) Severity

(b) TRIP10

(c) UMAP Dimension 1

(d) UMAP Dimension 2

Supplementary Figure 9. Additional analyses of the UMAP embedding for MS. (a) The UMAP dimensions did not correlate with MS severity as assessed by EDSS. However, lower ranked genes such as TRIP10 correlated with both dimensions in (b). We expanded the analysis to the top 30 genes and plot the genes with the highest correlations to UMAP dimension one and two in (c) and (d), respectively.

Proofs

Lemma 1. *Assume Lipschitz continuity of the conditional expectation for all $N \geq n_0$:*

$$\mathbb{E}\left|\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) - \mathbb{E}(Y|\boldsymbol{U}, L, B)\right| \leq \mathbb{E}C_N \left|\widetilde{\boldsymbol{U}} - \frac{\boldsymbol{U}}{dL}\right|, \tag{6}$$

where $d = \frac{\pi_{\boldsymbol{U}B}}{\sum_{\widetilde{X}_i \in \widetilde{A}} \widetilde{x}_i \pi_{iB}}$, $C_N \in O(1)$ is a positive constant, and we have taken an outer expectation on both sides. Then $\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) = \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, L, B)$ almost surely.

Proof. We can write the following sequence:

$$\begin{aligned}
&\mathbb{E}\left|\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) - \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, L, B)\right| = \mathbb{E} \lim_{N\to\infty} \left|\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) - \mathbb{E}(Y|\boldsymbol{U}, L, B)\right| \\
\leq\, &\mathbb{E} \lim_{N\to\infty} C_N \left|\widetilde{\boldsymbol{U}} - \frac{\boldsymbol{U}}{dL}\right| \leq C\mathbb{E}\left|\widetilde{\boldsymbol{U}} - \frac{1}{d} \lim_{N\to\infty} \frac{\boldsymbol{U}}{L}\right| = C\mathbb{E}\left|\widetilde{\boldsymbol{U}} - \frac{1}{d}\widetilde{\boldsymbol{U}}d\right| = 0,
\end{aligned}$$

where we have applied Expression (6) at the first inequality. We have $C_N \leq C$ for all $N \geq n_0$ in the second inequality because $C_N \in O(1)$. With the above bound, choose $a > 0$ and invoke the Markov inequality:

$$\mathbb{P}\left(\left|\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) - \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, L, B)\right| \geq a\right) \leq \frac{1}{a}\mathbb{E}\left|\mathbb{E}(Y|\widetilde{\boldsymbol{U}}) - \lim_{N\to\infty} \mathbb{E}(Y|\boldsymbol{U}, L, B)\right| = 0.$$

The conclusion follows because we chose a arbitrarily. □

Proposition 1. *If $E_i \not\perp\!\!\!\perp Y$ or $E_i \not\perp\!\!\!\perp Y|\text{Pa}(\widetilde{X}_i)$ (or both), then E_i is a root cause of Y.*

Proof. If $E_i \not\perp\!\!\!\perp Y$ or $E_i \not\perp\!\!\!\perp Y|\text{Pa}(\widetilde{X}_i)$ (or both), then E_i and Y are d-connected by the global Markov property. Since E_i is a root vertex, the d-connection implies that there exists a directed path from E_i to Y. □

Proposition 2. *We have* $\mathbb{P}(Y|E_i, \text{Pa}(\widetilde{X}_i)) = \mathbb{P}(Y|\widetilde{X}_i, \text{Pa}(\widetilde{X}_i))$ *under Equation* (3).

Proof. We can write:

$$\mathbb{P}(Y|E_i, \text{Pa}(\widetilde{X}_i)) = \mathbb{E}_{\widetilde{X}_i|E_i,\text{Pa}(\widetilde{X}_i)} \mathbb{P}(Y|E_i, \widetilde{X}_i, \text{Pa}(\widetilde{X}_i)) = \mathbb{P}(Y|E_i, \widetilde{X}_i, \text{Pa}(\widetilde{X}_i)) = \mathbb{P}(Y|\widetilde{X}_i, \text{Pa}(\widetilde{X}_i)).$$

The second equality follows because $\widetilde{X}_i$ is a constant given E_i and $\text{Pa}(\widetilde{X}_i)$. The third equality follows by the global Markov property because Y is a terminal vertex. □

Theorem 2. *(Fisher consistency) Consider the same assumption as Lemma 1. If unconditional d-separation faithfulness holds, then RCSP recovers* Φ *almost surely as* $N \to \infty$.

Proof. If $X_k \not\perp\!\!\!\perp P_i$ in Line 2 of Algorithm 1, then X_k is a descendant of the root vertex P_i under the global Markov property. Similarly, if X_k is a descendant of P_i, then X_k is d-connected to P_i so $X_k \not\perp\!\!\!\perp P_i$ by unconditional d-separation faithfulness. Hence, $\text{SD}(\widetilde{X}_i)$ contains only and all the surrogate descendants of $\widetilde{X}_i$ for each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$. This in turn implies that $\text{SA}(\widetilde{X}_i)$ in Line 5 of Algorithm 1 contains only and all the surrogate ancestors of $\widetilde{X}_i$. Hence, RCSP now has access to the correct set $\text{SA}(\widetilde{X}_i)$ as well as B for each $\widetilde{X}_i \in \widetilde{\boldsymbol{X}}$. We finally invoke Theorem 1 to conclude that RCSP recovers Φ almost surely as $N \to \infty$. □

Supplementary Materials References

1. Olsen, T. W. & Feng, X. The minnesota grading system of eye bank eyes for age-related macular degeneration. *Investig. Ophthalmol. Vis. Sci.* **45**, 4484–4490 (2004).
2. Orian, J. M. *et al.* Platelets in multiple sclerosis: early and central mediators of inflammation and neurodegeneration and attractive targets for molecular imaging and site-directed therapy. *Front. Immunol.* **12**, 620963 (2021).
3. Basile, M. S., Bramanti, P. & Mazzon, E. The role of cytotoxic t-lymphocyte antigen 4 in the pathogenesis of multiple sclerosis. *Genes* **13**, 1319 (2022).