

Discovering Root Causal Genes with High Throughput Perturbations

Eric V Strobl , Eric R Gamazon

University of Pittsburgh, Pittsburgh, United States • Vanderbilt University Medical Center, Nashville, United States

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Reviewed Preprint

v2 • February 10, 2025

Revised by authors

Reviewed Preprint

v1 • October 15, 2024

eLife Assessment

This work provides an **important** framework for understanding the primary causes of disease. While the theoretical results rely on strong assumptions about the underlying causal mechanisms, the authors provide **solid** empirical evidence that the framework is robust to modest violations of these assumptions.

<https://doi.org/10.7554/eLife.100949.2.sa3>

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.

Introduction

Root causes of disease correspond to the most upstream causes of a diagnosis with strong causal effects on the diagnosis. *Pathogenesis* refers to the causal cascade from root causes to the diagnosis. Genetic and non-genetic factors may act as root causes and affect gene expression as an intermediate step during pathogenesis. We introduce root causal gene expression levels – or *root causal genes* for short – that correspond to the initial changes to *gene expression* induced by genetic and non-genetic root causes that have large causal effects on a downstream diagnosis (**Figure 1 (a)** ). Root causal genes differ from core genes that directly cause the diagnosis and

thus lie at the end, rather than at the beginning, of pathogenesis (Boyle et al., 2017 [↗](#)). Root causal genes also generalize (the expression levels of) driver genes that only account for the effects of somatic mutations primarily in cancer (Martínez-Jiménez et al., 2020 [↗](#)).

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 (Nagral, 2014). 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 (Cano-Gamez and Trynka, 2020 [↗](#)). Complex diseases often have an overwhelming number of causes but, just like a machine usually breaks down due to one or a few root causal problems, the root causal genes may only represent a small subset of the genes because the causal effects of only a few root causes are large (Figure 1 (b) [↗](#)). 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 or predictive relations, sometimes visually represented as gene regulatory networks (Costa-Silva et al., 2017 [↗](#); Ellington et al., 2023 [↗](#)). Other methods even identify causal relations (Friedman et al., 2000 [↗](#); Wang et al., 2023 [↗](#); Wen et al., 2023 [↗](#); Buschur et al., 2020 [↗](#)), but none pinpoint the *first* gene expression levels that ultimately generate the vast majority of pathogenesis. Simply learning a causal graph does not resolve the issue because causal graphs do not summarize the effects of *unobserved* root causes, such as unmeasured environmental changes or variants, that are needed to identify all root causal genes. We therefore define the Root Causal Strength (RCS) score to identify all 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 (Dixit et al., 2016 [↗](#); Adamson et al., 2016 [↗](#); Datlinger et al., 2017 [↗](#)). Experiments demonstrate marked improvements in performance, when investigators have access to a large bulk RNA-seq dataset and a genomewide Perturb-seq dataset from a cell line of a disease-relevant tissue. Finally, application of the algorithm to two complex diseases with disparate pathogeneses recovers an *omnigenic root causal model*, where a small set of root causal genes drive pathogenesis but impact many downstream genes within each patient. As a result, nearly all gene expression levels are correlated with the diagnosis at the population level.

Results

We briefly summarize the Methods in the first two subsections.

Definitions

Differential expression analysis identifies differences in gene expression levels between groups Y (Costa-Silva et al., 2017 [↗](#)). 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

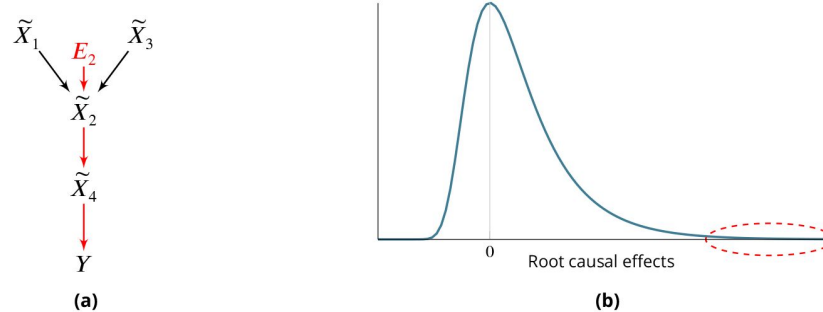


Figure 1.

(a) Toy example where a variable E_2 simultaneously models genetic and non-genetic root causes that jointly have a large causal effect on a diagnose Y through gene expression $\tilde{X}$. E_2 first affects the gene expression level $\tilde{X}_2$, or the root causal gene. The root causal gene then affects other downstream levels during pathogenesis, including the core (or direct causal) gene $\tilde{X}_4$, to ultimately induce a diagnosis Y . (b) We hypothesize that the causal effects of most root causes are small, but a few are large (red ellipse), in each patient with disease. As a result, the distribution of these *root causal effects* tends to be right skewed in disease.

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 2 (a)** – where p vertices $\tilde{X}$ represent true gene expression levels in a bulk sample and Y denotes the patient symptoms or diagnosis. The set $\tilde{X}$ 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 $\tilde{X}$. The set $\text{Pa}(\tilde{X}_i)$ refers to the *parents* of $\tilde{X}_i \in \tilde{X}$, or those variables with an edge directed into $\tilde{X}_i$. For example, $\text{Pa}(\tilde{X}_2) = \{\tilde{X}_1, \tilde{X}_3\}$ in **Figure 2 (a)**. A *root vertex* corresponds to a vertex with no parents.

We can associate $\mathbb{G}$ with the structural equation $\tilde{X}_i = f_i(\text{Pa}(\tilde{X}_i), E_i)$ for each $\tilde{X}_i \in \tilde{X}$ that links each vertex to its parents and error term E_i (Pearl, 2009). The error term E_i is not simply a regression residual but instead represents the conglomeration of unobserved explanatory variables that only influence $\tilde{X}_i$, such as unobserved transcriptional regulators, certain genetic variants and specific environmental conditions. We thus also include the error terms E in the directed graph of **Figure 2 (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 $\tilde{X}_i$ on Y as the following absolute difference (**Figure 2 (c)**):

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

We prove the last equality in the Methods. As a result, RCS Φ_i directly measures the contribution of the gene $\tilde{X}_i$ on Y according to its error term E_i without recovering the error term values. The algorithm does not impose distributional assumptions or 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 $\text{Pa}(\tilde{X}_i)$ and $\tilde{X}_i$ may differ between patients. We have $\Phi_i = 0$ when E_i is not a cause of Y , and we say that the gene $\tilde{X}_i$ is a *patient-specific root causal gene* if $\Phi_i \gg 0$, or its (conditional) root causal effect is large as depicted by the red ellipse in **Figure 1 (b)**.

Algorithm

We propose an algorithm called Root Causal Strength using Perturbations (RCSP) that estimates $\Phi = \{\Phi_1, \dots, \Phi_p\}$ from genes measured in both bulk RNA-seq and Perturb-seq datasets derived from possibly independent studies but from the same tissue type. We rely on bulk RNA-seq instead of single cell RNA-seq in order to obtain many samples of the label Y . We focus on statistical estimation rather than statistical inference because $\Phi_i > 0$ when E_i causes Y under mild conditions, so we reject the null hypothesis that $\Phi_i = 0$ for many genes if many gene expression levels cause Y . However, just like a machine typically breaks down due to only one or a few root causal problems, we hypothesize that only a few genes have large RCS scores $\Phi_i \gg 0$ even in complex disease.

Estimating Φ requires access to the true gene expression levels $\tilde{X}$ and the removal of the effects of confounding. We first control for batch effects representing unwanted sources of technical variation such as different sequencing platforms or protocols. We however can only obtain imperfect counts X from RNA sequencing even within each batch (**Figure 2 (d)**). Measurement error introduces confounding as well because it prevents us from exactly controlling for the causal effects of the gene expression levels. Investigators usually mitigate measurement error by normalizing the gene expression levels by sequencing depth. We show in the Methods that the Poisson distribution approximates the measurement error distribution induced by the sequencing process to high accuracy (Choudhary and Satija, 2022; Sarkar and Stephens, 2021). We

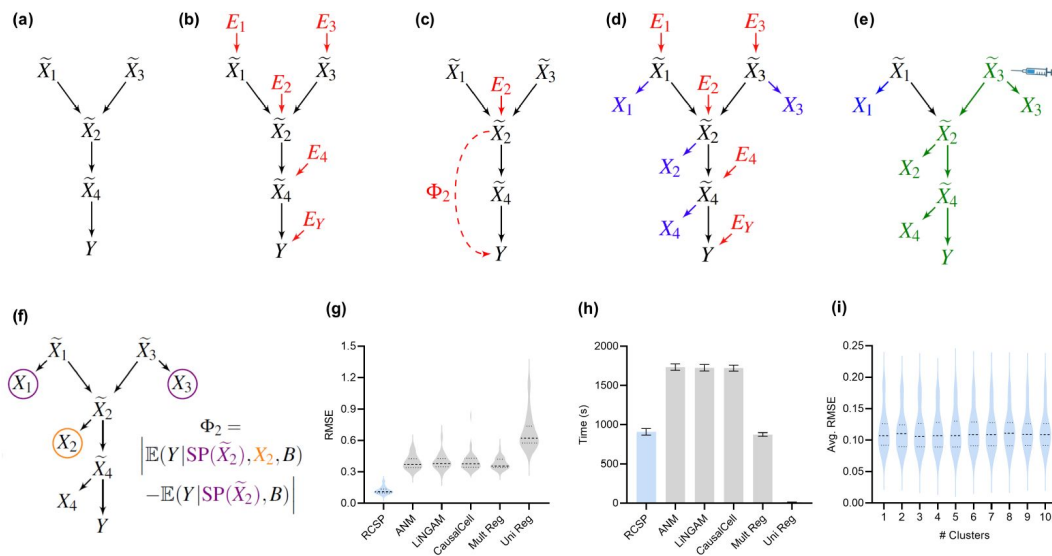


Figure 2.

Method overview and synthetic data results. (a) We consider a latent causal graph over the true counts $\tilde{\mathbf{X}}$. (b) We augment the graph with error terms $\mathbf{E}$ such that each $E_i \in \mathbf{E}$ in red has an edge directed towards $\tilde{X}_i \in \tilde{\mathbf{X}}$. (c) The RCS of $\tilde{X}_2$, denoted by Φ_2 , quantifies the magnitude of the *conditional root causal effect*, or the strength of the causal effect from E_2 to Y conditional on $\text{Pa}(\tilde{X}_2)$. (d) We cannot observe $\tilde{\mathbf{X}}$ in practice but instead observe the noisy surrogates $\mathbf{X}$ in blue corrupted by Poisson measurement error. (e) Perturbing a variable such as $\tilde{X}_3$ changes the marginal distributions of downstream variables shown in green under mild conditions. (f) RCSP thus uses the perturbation data to identify (an appropriate superset of) the surrogate parents for each variable in order to compute Φ . (g) Violin plots show that RCSP achieved the smallest RMSE to the ground truth RCS values in the synthetic data. (h) 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. (i) Finally, RCSP maintained low RMSE values regardless of the number of clusters considered.

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:

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

where $\text{SP}(\tilde{X}_i)$ refers to the *surrogate parents* of $\tilde{X}_i$, or the variables in $\mathbf{X}$ associated with $\text{Pa}(\tilde{X}_i) \subseteq \tilde{\mathbf{X}}$. 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 (**Figures 2 (e)** [↗](#) and **(f)** [↗](#)). 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; we empirically verify the robustness in the Supplementary Materials. RCSP finally performs the two non-linear regressions needed to estimate $\mathbb{E}(Y | \text{SP}(\tilde{X}_i), X_i, B)$ and $\mathbb{E}(Y | \text{SP}(\tilde{X}_i), B)$ 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 $\text{SP}(\tilde{X}_i)$.

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) (Peters et al., 2014 [↗](#); Strobl and Lasko, 2023a), the Linear Non-Gaussian Acyclic Model (LiNGAM) (Peters et al., 2014 [↗](#); Strobl and Lasko, 2022 [↗](#)), CausalCell (Wen et al., 2023 [↗](#)), 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 2 (g)** [↗](#) 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 RCSP against 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 2 (h)** [↗](#)). RCSP maintained the lowest RMSE even in the cyclic case, and the performance of the algorithm remained robust to differences between the directed graphs underlying the bulk RNA-seq and Perturb-seq data (Supplementary Materials). We conclude that RCSP both scalably and accurately estimates Φ .

We will cluster the RCS values in real data to find patient subgroups. We therefore also performed hierarchical clustering using Ward's method (Ward Jr, 1963 [↗](#)) on the values of Φ estimated by RCSP with the synthetic data. We then computed the RMSEs and averaged them within each

cluster. We found that RCSP maintained low average RMSE values regardless of the number of clusters considered (**Figure 2 (i)**). 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) and a Perturb-seq dataset of 247,914 cells generated from an immortalized retinal pigment epithelial (RPE) cell line (Ratnapriya et al., 2019; Replogle et al., 2022). The Perturbseq dataset contains knockdown experiments of 2,077 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 (Hadziahmetovic and Malek, 2021), so age is a known root cause of the disease. We therefore determined if RCSP identified age as a root cause. Note that RCSP does not need perturbation data of age to compute the RCS values of age, since age has no parents in the directed graph. The algorithm estimated a heavy tailed distribution of the RCS values indicating that most of the RCS values deviated away from zero (**Figure 3 (a)**). The Deviation of the RCS (D-RCS), or the standard deviation from an RCS value of zero, measures the tailedness of the distribution while preserving the unit of measurement. The D-RCS of age corresponded to 0.46 – more than double that of the nearest gene (**Figure 3 (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 3 (b)**. Notice that the histogram of D-RCS scores in **Figure 3 (b)** mimics a folded distribution of **Figure 1 (b)**. Thus, few D-RCS scores had large values implying the existence of only a few 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 (Barouch and Miller, 2007). We therefore expect the root causal genes of AMD to include genes involved in oxidative stress during early pathogenesis. The gene MIPEP with the highest D-RCS score in **Figure 3 (d)** indeed promotes the maturation of oxidative phosphorylation-related proteins (Shi et al., 2011). The second gene SLC7A5 is a solute carrier that activates mTORC1 whose hyperactivation increases oxidative stress via lipid peroxidation (Nachev et al., 2021; Go et al., 2020). The gene HEATR1 is involved in ribosome biogenesis that is downregulated by oxidative stress (Turi et al., 2018). The top genes discovered by RCSP thus identify pathways known to be involved in oxidative stress. We further verified that measurement error did not explain their large D-RCS scores in Supplementary Materials.

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 3 (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 (Nicklin et al., 2009);

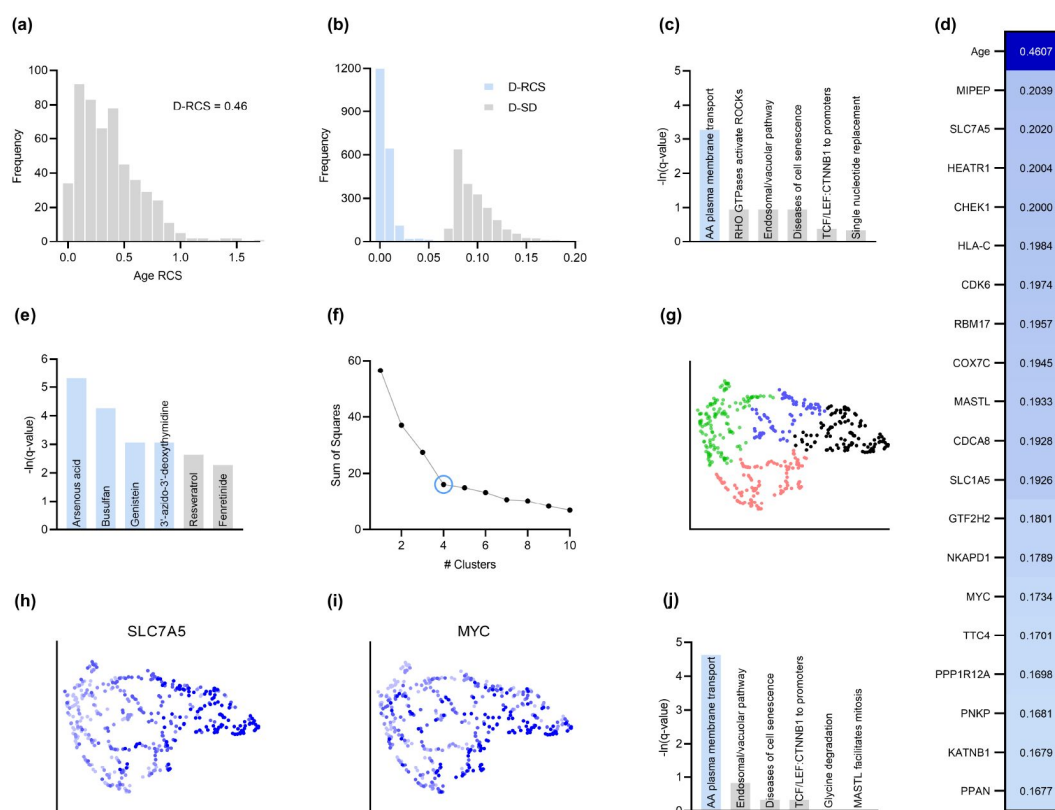


Figure 3.

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).

Beaumat et al., 2019). 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 (Nache et al., 2021; Go et al., 2020). 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 (Su et al., 2020), but the other three significant drugs have therapeutic potential (Figure 3 (e)). Busulfan decreases the requirement for intravitreal anti-VEGF injections (Dalvin et al., 2022). Genistein is a protein kinase inhibitor that similarly attenuates neovascularization (Kinoshita et al., 2014) and blunts the effect of ischemia on the retina (Kamalden et al., 2011). Finally, a metabolite of the antiviral agent 3'-azido-3'-deoxythymidine inhibits neovascularization and mitigates RPE degeneration (Narendran et al., 2020). 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 by the elbow method on the sum of squares plot (Figures 3 (f) and (g), respectively). The RCS scores of 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 3 (h). We found three exceptions to this rule among the top 30 genes (example in Figure 3 (i) and see Supplementary Materials). RCSP thus detected genes with large RCS scores primarily in the black cluster of Figure 3 (g). Pathway enrichment analysis within this cluster alone yielded supra-significant results on the same pathway detected in the global analysis (Figure 3 (j) versus Figure 3 (c)). Furthermore, drug enrichment analysis results by cluster 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 undifferentiated blast cells that can be induced to differentiate into lymphoblasts, or the precursors of T cells and other lymphocytes (Kim et al., 2021; Replogle et al., 2022). 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 4 (a)). The distribution of gene D-RCS scores concentrated around zero with a long tail, whereas the distribution of gene D-SD scores concentrated around a relatively larger value of 0.3 (Figure 4 (b)). RCSP thus detected an omnigenic root causal model with a few root causal genes but many correlated genes.

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 (Fletcher et al., 2010). Damage induced by the T cells also perturbs cellular homeostasis and leads to the accumulation of misfolded proteins (Andhavarapu et al., 2019). 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 4 (d)). MNT is a MYC antagonist that modulates the proliferative and pro-survival signals of T cells after engagement of the T cell receptor (Gnanaprakasam and Wang, 2017). Similarly, CERCAM is an adhesion molecule expressed at high levels in microvessels of the brain that increases leukocyte

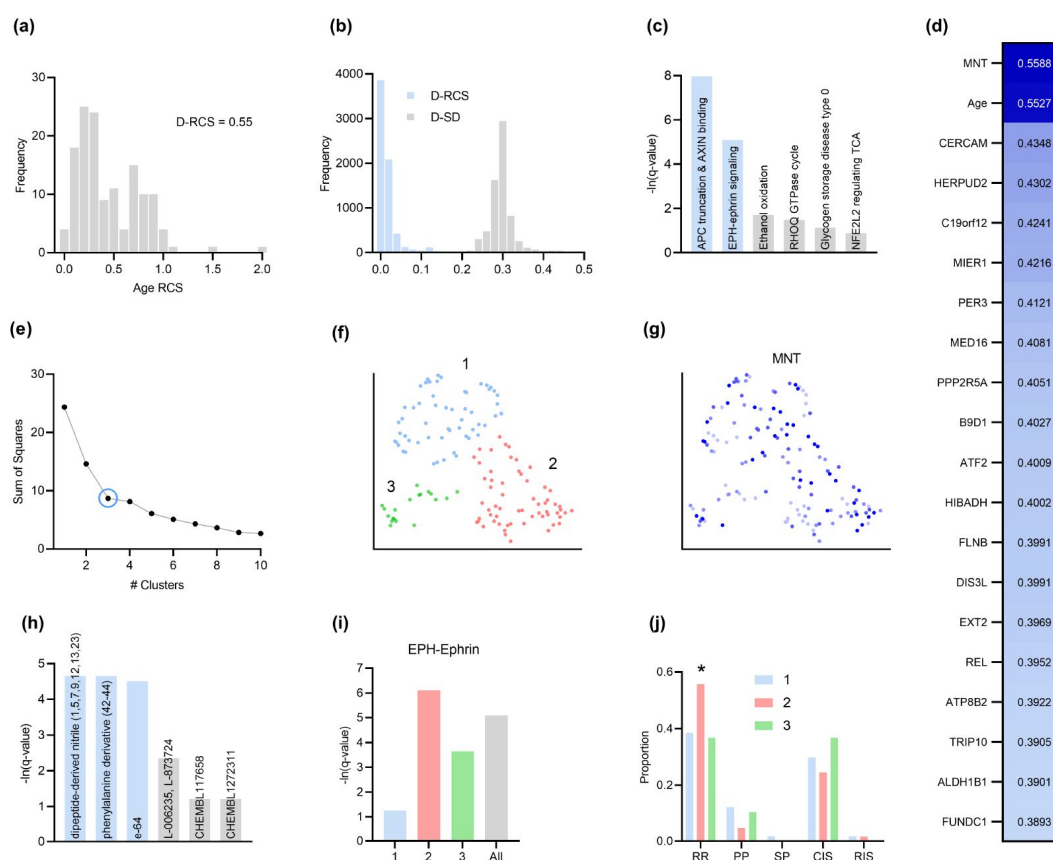


Figure 4.

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).

transmigration across the blood brain barrier (Starzyk et al., 2000 [↗](#)). HERPUD2 is involved in the endoplasmic-reticulum associated degradation of unfolded proteins (Kokame et al., 2000 [↗](#)). 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 4 (d) [↗](#) versus Figure 3 (d) [↗](#)). Measurement error did not account for the high scores (Supplementary Materials). 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 4 (c) [↗](#)). APC and AXIN are both members of the Wnt signaling pathway and regulate levels of beta-catenin (Spink et al., 2000 [↗](#)). Furthermore, inhibition of Wnt/beta-catenin causes CD4+ T cell infiltration into the central nervous system via the blood brain barrier in MS (Lengfeld et al., 2017 [↗](#)). Ephrins similarly regulate T cell migration into the central nervous system (Luo et al., 2016 [↗](#)) and are overexpressed in MS lesions (Sobel, 2005 [↗](#)). 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 4 (e) [↗](#) revealed the presence of three clusters by the elbow method. We plot the three clusters in a UMAP embedding in Figure 4 (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 4 (d) [↗](#); we plot the MNT gene as an example in Figure 4 (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 4 (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, phenylalanine derivatives, e-64, L-006235 and L-873724 (Figure 4 (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 multiple experimental findings highlighting the therapeutic efficacy of cathepsin inhibitors in a subgroup of MS patients responsive to interferon therapy (Haves-Zburuf et al., 2011 [↗](#); Burster et al., 2007 [↗](#)).

Prior research has also shown that EPH-ephrin signaling is more prevalent in relapsing-remitting multiple sclerosis than in other subtypes of the disease (Golan et al., 2021 [↗](#)). EPH-ephrin signaling survived FDR correction in our analysis (Figure 4 (c) [↗](#)). Furthermore, the pathway was more enriched in the pink cluster than in the other two (Figure 4 (i) [↗](#)). The pink cluster indeed contained a higher proportion of patients with the relapsing-remitting subtype (Figure 4 (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, or the gene expression levels directly regulated by root causes with large causal effects on Y, by modeling the root causes 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 directly cause

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, or the magnitude of the conditional causal effect of each error term, which we can compute using gene expression levels alone. 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 only a few 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, but virtually all genes were correlated with Y . This omnigenic model, where “omni-” refers to the nearly all genes correlated with Y , differs from the omnigenic model involving *core genes* (Boyle et al., 2017 [↗](#)). Boyle et al. define core genes as genes that directly affect disease risk. The authors 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. In contrast, root causal genes may not directly cause Y but lie substantially upstream of Y in the causal graph. The error terms of upstream root causal genes affect many downstream genes that include both ancestors and non-ancestors of Y (Figure 5 [↗](#)). These downstream genes contain traces of the root causal gene error terms that induce the many correlations with Y . The root causal model thus assumes sparsity in upstream root causal genes, whereas the core gene model assumes sparsity in the downstream direct causal genes (The omnigenic root causal model makes no statement about the number of direct causal genes, so direct causal genes may be sparse or dense). Further, each causal genetic variant tends to have only a small effect on disease risk in complex disease because the variant can directly cause Y or directly cause any causal gene including those with small root causal effects on Y ; thus, all error terms that cause Y can model genetic effects on Y . However, the root causal model further elaborates that genetic and non-genetic factors often combine to produce a few root causal genes with large root causal effects, where non-genetic factors typically account for the majority of the large effects in complex disease. Many variants may therefore cause many genes in diseases with only a few root causal genes. We finally emphasize that the root causal model accounts for all deleterious effects of the root causal genes, whereas the core gene model only captures the deleterious effects captured by the diagnosis Y . For example, the *disease* of diabetes causes retinopathy, but retinopathy is not a part of the diagnostic criteria of diabetes. As a result, the gene expression levels that cause retinopathy but not the *diagnosis* of diabetes are not core genes, even though they are affected by the root causal genes. The sparsity of the root causal genes, the focus on the combined effects of genetic and non-genetic root causes, and the ability to account for root causal effects not represented by the target 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 (Strobl and Lasko, 2022 [↗](#); Strobl et al., 2024 [↗](#); Strobl and Lasko, 2023a). 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 (Sarkar and Stephens, 2021 [↗](#)). 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 (Strobl and Lasko, 2023b). 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.

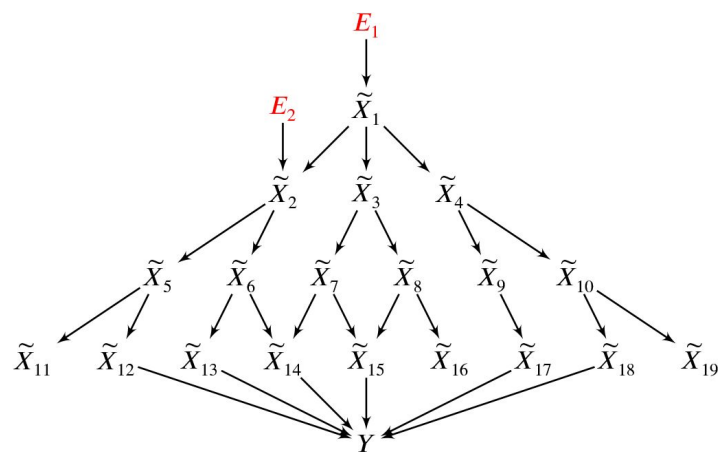


Figure 5.

In this example, two root causal genes $\tilde{X}_1$ and $\tilde{X}_2$ affect many downstream genes and ultimately cause Y . Thus all genes $\tilde{X}_1, \dots, \tilde{X}_{19}$ correlate with Y , but only $\tilde{X}_1$ and $\tilde{X}_2$ have large root causal effects on Y . The omnigenic root causal model posits that only a few root causal genes affect many downstream genes, so that nearly all genes are correlated with Y . Causal genetic variants can directly cause Y or cause any gene expression level that causes Y – including those with small root causal effects – but only $\tilde{X}_1$ and $\tilde{X}_2$ have large root causal effects on Y due to genetic *and non-genetic* root causes modeled by E_1 and E_2 . In contrast, the core gene model assumes only a few direct causal genes $\tilde{X}_{12}, \tilde{X}_{13}, \tilde{X}_{14}, \tilde{X}_{15}, \tilde{X}_{17}, \tilde{X}_{18}$. These core genes do not account for the deleterious causal effects of E_1 and E_2 on $\tilde{X}_{11}, \tilde{X}_{16}$ and $\tilde{X}_{19}$.

This study carries other limitations worthy of addressing in future work. 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 signed RCS (or expected conditional root causal effect) do not differ by much in practice (Supplementary Materials), but future work may focus on exactly identifying both the strength and direction of the unconditional causal effects of the error terms. Furthermore, RCS achieves patient but not cell-specificity because the algorithm relies on phenotypic labels obtained from bulk RNAseq. RCSP thus cannot identify the potentially different root causal genes present within distinct cell populations. Modern genome-wide Perturb-seq datasets also adequately perturb and measure only a few thousand, rather than all, gene expression levels. RCSP can only identify root causal genes within this perturbed and measured subset. Fourth, RCSP accounts for known batch effects and measurement error but cannot adjust for unknown confounding. Finally, RCSP assumes a directed acyclic graph. We can transform a directed graph with cycles into an acyclic one under equilibrium, but real biological distributions vary across time (Spirtes, 1995 [↗](#); Bongers et al., 2021 [↗](#)). Future work should thus aim to estimate cell-specific root causal effects under latent confounding and time-varying distributions.

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 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 gene discovery.

Methods and Materials

Background on Causal Discovery

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

$$Z_i = f_i(\text{Pa}(Z_i), E_i), \quad \forall Z_i \in \mathbf{Z} \quad (3)$$

where f_i is a function of the *parents*, or direct causes, of Z_i and an error term $E_i \in \mathbf{E}$. The error terms $\mathbf{E}$ 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* or *sink vertex* is not a parent of any other vertex.

We can associate a directed graph to $\mathbf{Z}$ by drawing a directed edge from each member of $\text{Pa}(Z_i)$ to Z_i for all $Z_i \in \mathbf{Z}$. 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 $\mathbf{E}$ and drawing a directed edge from each $E_i \in \mathbf{E}$ 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 $\mathbf{W} \subseteq \mathbf{Z} \setminus \{Z_i, Z_j\}$ 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 $\mathbf{W}$ and no non-collider is in $\mathbf{W}$. 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 $\mathbf{W}$ if they are d-separated given $\mathbf{W}$. 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 $\mathbf{W} = \emptyset$.

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 snapshot of a biological causal process using an SEM over $\tilde{\mathbf{X}} \cup \mathbf{Y}$ 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 $\tilde{X}_i \in \tilde{\mathbf{X}}$ corresponds to the total number of RNA molecules of a unique gene in a single cell or bulk tissue sample. The error terms model root causes that are outside of gene expression, such as genetic variation or environmental factors. Moreover, the relation from gene expression to Y is stochastic because $Y = f_Y(\text{Pa}(Y), E_Y)$, where E_Y introduces the stochasticity. Two individuals may therefore have the exact same error term values over $\tilde{\mathbf{X}}$ but different instantiations of Y .

We unfortunately cannot observe $\tilde{\mathbf{X}}$ in practice but instead measure a corrupted count $\mathbf{X}$ 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 $\tilde{X}_i \in \tilde{\mathbf{X}}$ within a sample at unequal coverage. Let π_{ij} denote the probability of mapping one molecule of $\tilde{X}_i$ in batch j so that $\sum_{i=1}^p \pi_{ij}$ is near zero. The law of rare events (Papoulis, 1984) implies that the Poisson distribution wellapproximates the library size N so that $N \sim \text{Pois}(\sum_{i=1}^p \tilde{X}_i \pi_{ij})$.

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

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

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

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

We conclude that the measurement error distribution follows a Poisson distribution to high accuracy. Multiple experimental results already corroborate this theoretical conclusion (Grün et al., 2014; Sarkar and Stephens, 2021; Choudhary and Satija, 2022).

We can represent the biology and the RNA sequencing in a single DAG over $\mathbf{X} \cup \tilde{\mathbf{X}} \cup \mathbf{B} \cup \mathbf{Y}$, where $\mathbf{B}$ denotes the batch, and Y the target variable representing patient symptoms or diagnosis. We provide a toy example in Figure 6. We draw $\mathbb{G}$ over $\mathbf{Z}$ in black and make each $\tilde{X}_i \in \tilde{\mathbf{X}}$ a

parent of $X_i \in \mathbf{X}$ in blue. We then include the root vertex B as a parent of all members of $\mathbf{X}$ in green. We augment this graph with the error terms of $\tilde{\mathbf{X}}$ in red and henceforth refer to the augmented DAG as $\mathbb{G}$. Repeated draws from the represented causal process generates a dataset.

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 $\tilde{\mathbf{X}}$ 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 $\tilde{\mathbf{A}}$ denote the set of housekeeping genes where $\tilde{X}_i = \tilde{x}_i$ is a constant for each $\tilde{X}_i \in \tilde{\mathbf{A}}$ similarly $\mathbf{A}$ refers to the corresponding set with Poisson measurement error. Let $N = n$ be large enough such that $\sum_{X_i \in \mathbf{A}} x_i > 0$ for each sample. Then dividing by $L \triangleq \sum_{X_i \in \mathbf{A}} X_i$ controls for sequencing depth in the following sense:

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{X_i}{\sum_{X_i \in \mathbf{A}} X_i} &= \lim_{N \rightarrow \infty} \frac{X_i/N}{\sum_{X_i \in \mathbf{A}} X_i/N} = \frac{P_{ij}}{\sum_{X_i \in \mathbf{A}} P_{ij}} \\ &= \frac{\tilde{X}_i \pi_{ij} / \sum_{i=1}^p \tilde{X}_i \pi_{ij}}{\sum_{\tilde{X}_i \in \tilde{\mathbf{A}}} \tilde{X}_i \pi_{ij} / \sum_{i=1}^p \tilde{X}_i \pi_{ij}} = \frac{\tilde{X}_i \pi_{ij}}{\sum_{\tilde{X}_i \in \tilde{\mathbf{A}}} \tilde{X}_i \pi_{ij}}, \end{aligned}$$

where we have divided $\tilde{X}_i \pi_{ij}$ by a constant in the last term. Thus, dividing by L removes measurement error within each batch as $N \rightarrow \infty$. 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 $\mathbb{E}(Y | \tilde{\mathbf{U}}, B) = \mathbb{E}(Y | \mathbf{U}, L, B)$ for any $\tilde{\mathbf{U}} \subseteq \tilde{\mathbf{X}}$ as $N \rightarrow \infty$. The equality holds almost surely under a mild smoothness condition:

Lemma 1.

Assume Lipschitz continuity of the conditional expectation for all $N \geq n_0$:

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

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

We delegate proofs to the Supplementary Materials unless proven here in the Methods. Note that $\lim_{N \rightarrow \infty} \frac{\mathbf{U}}{dL} = \tilde{\mathbf{U}}$, so the Lipschitz assumption intuitively means that accurate estimation of $\tilde{\mathbf{U}}$ implies accurate estimation of $\mathbb{E}(Y | \tilde{\mathbf{U}})$. 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.

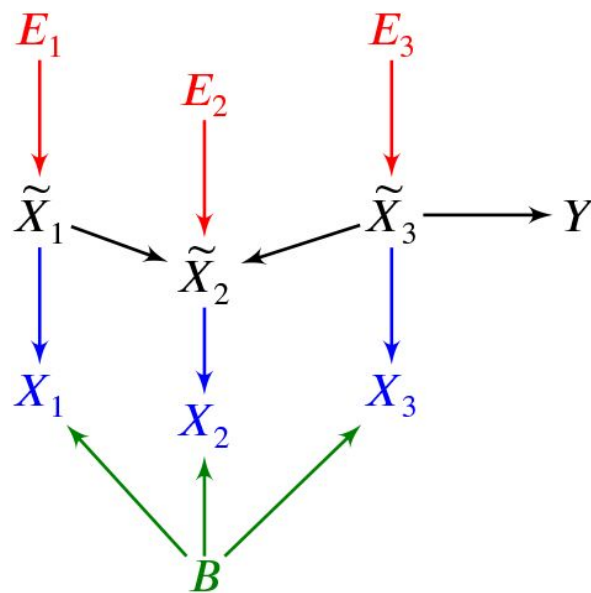


Figure 6.

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

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 $Y \perp\!\!\!\perp L \mid (U, B)$ for any N , so that $\mathbb{E}(Y \mid \tilde{U}) = \lim_{N \rightarrow \infty} \mathbb{E}(Y \mid U, L, B) = \lim_{N \rightarrow \infty} \mathbb{E}(Y \mid U, B)$ almost surely. We have proved:

Theorem 1.

Consider the same assumption as Lemma 1. Then $\mathbb{E}(Y \mid \tilde{U}) = \lim_{N \rightarrow \infty} \mathbb{E}(Y \mid U, B)$ 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 $E_i \in \mathbf{E}$ on Y as $Y_i \triangleq \mathbb{P}(Y \mid E_i) - \mathbb{P}(Y)$ (Strobl, 2024 [\[1\]](#); Strobl and Lasko, 2023c).

We can identify root causes using the following result:

Proposition 1.

If $E_i \not\perp\!\!\!\perp Y \mid Y$ or $E_i \not\perp\!\!\!\perp Y \mid \text{Pa}(\tilde{X}_i)$ (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 (Strobl, 2022 [\[2\]](#)).

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

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

The above quantity corresponds to the *conditional root causal effect* but not the root causal effect Y_i due to the extra conditioning on $\text{Pa}(\tilde{X}_i)$. The two terms may also differ in direction; if $\Delta_i > 0$, then this does not imply that $Y_i > 0$, and similarly for negative values. The two variables thus represent different quantities but – in terms of priority – we would estimate Y_i when we have nonzero Δ_i . Experimental results indicate that Y_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 Y_i and Δ_i depend on the unknown error term E_i . We can however substitute E_i with $\tilde{X}_i$ in Δ_i due to the following result:

Proposition 2.

We have $\mathbb{P}(Y \mid E_i, \text{Pa}(\tilde{X}_i)) = \mathbb{P}(Y \mid \tilde{X}_i, \text{Pa}(\tilde{X}_i))$ under Equation (3).

We can thus compute the conditional root causal effect Δ_i without access to the error terms:

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

We can determine the root causal status of E_i on Y when $\Delta_i \neq 0$ per Proposition 1. Nevertheless, the term “root cause” in colloquial language refers to two concepts simultaneously: a root vertex that causes Y and has a large causal effect on Y . We thus say that $\tilde{X}_i$ is a *root causal gene* of Y if $\Delta_i \gg 0$.

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:

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

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 $\text{SP}(\tilde{X}_i) \subseteq X$ denotes the *surrogate parents* of $\tilde{X}_i$ corresponding to the variables in X associated with $\text{Pa}(\tilde{X}_i) \subseteq \tilde{X}$. The last equality holds almost surely as $N \rightarrow \infty$ by Theorem 1.

We call $\Phi \triangleq |\Gamma|$ the *Root Causal Strength* (RCS) of $\tilde{X}_i$ on Y . The RCS obtains a unique value $\Phi = \phi_{ij}$ for each patient j . We say that $\tilde{X}_i$ is a root causal gene of Y for patient if $\phi_{ij} \gg 0$, since we posit a right skewed distribution of conditional root causal effects for each patient as in **Figure 1 (b)**. 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 $\tilde{X}_i$ 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 much larger than zero, respectively. Note that we do not specify a particular cutoff value for large (conditional) root causal effects, since the root causal effects likely lie on a continuous graduated scale as opposed to approximately two binary values. Nevertheless, visual inspection of the RCS or D-RCS histograms in disease should approximate a power law, where a large mass is concentrated around zero and a long tail extends to the right similar to folding **Figure 1 (b)**.

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 $\tilde{X}_i \in \tilde{X}$ 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 (Colombo et al., 2014) and require restrictive assumptions such as d-separation faithfulness (Spirtes et al., 2000) or specific functional relations (Peters et al., 2014).

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:

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

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

We discover the surrogate ancestors using unconditional independence tests. For any $X_k \in \mathbf{X}$, 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 reject the null and 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 $SD(\tilde{X}_i)$. Curating every $X_j \in \mathbf{X}$ such that $X_j \in SD(\tilde{X}_i)$ into $SA(\tilde{X}_i)$ yields the surrogate ancestors of $\tilde{X}_i$ as desired.

Procedure

We now introduce an algorithm called Root Causal Strength using Perturbations (RCSP) that discovers the surrogate ancestors of each variable $\tilde{X}$ 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 $\tilde{\mathbf{X}}$ 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 $X_i \in \mathbf{X}$ from the bulk RNA-seq in Line 6. The algorithm thus outputs the RCS scores Φ as desired.

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 $N \rightarrow \infty$.

Algorithm 1

Root Causal Strength using Perturbations (RCSP)

Input: bulk RNA-seq data with batches B , Perturb-seq data

Output: RCS scores Φ

```

1: for each  $X_i \in \mathbf{X}$  do
2:    $SD(\tilde{X}_i) \leftarrow$  all  $X_k \in \mathbf{X}$  s.t.  $X_k \not\perp\!\!\!\perp P_i$  in Perturb-seq
3: end for
4: for each  $X_i \in \mathbf{X}$  do
5:    $SA(\tilde{X}_i) \leftarrow$  all  $X_k \in \mathbf{X}$  s.t.  $X_i \in SD(\tilde{X}_k)$ 
6:   Compute  $\Phi_i$  using Eq. (5) in bulk RNA-seq
7: end for
```

We engineered RCSP to only require *unconditional* d-separation faithfulness because real distributions may not obey full d-separation faithfulness (Strobl, 2022 [4](#)).

Synthetic Data

Simulations

We generated a linear SEM obeying Equation (3) specifically as $\tilde{X}_i = \tilde{X}\beta_i + E_i$ for every $\tilde{X}_i \in \tilde{X}$ and similarly $Y = \tilde{X}\beta_Y + E_Y$. We included $p + 1 = 2500$ variables in $\tilde{X} \cup Y$. 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 $\tilde{X}_i \leftarrow \text{softplus}(\tilde{X}_i)$ and then obtained the corrupted surrogate X_i distributed $\text{Pois}(\tilde{X}_i \pi_{ij})$ for each $\tilde{X}_i \in \tilde{X}$ and batch j . We chose Y uniformly at random from the set of vertices with at least one parent and no children. We drew 200 samples for the bulk RNA-seq data to mimic a large but common dataset size. We introduced knockdown perturbations in Perturb-seq by subtracting an offset of two in the softplus function: $\tilde{X}_i \leftarrow \text{softplus}(\tilde{X}_i - 2)$. We finally drew 200 samples for the control and each perturbation condition to generate the Perturb-seq data. We repeated the above procedure 30 times.

Comparators

We compared RCSP against the following four algorithms:

1. Additive noise model (ANM) (Peters et al., 2014; Strobl and Lasko, 2023a): performs nonlinear regression of X_i on $\text{Pa}(X_i) \cup B$ and then regresses Y on the residuals $E \setminus E_i$ to estimate $|\mathbb{E}(Y | E \setminus E_i) - \mathbb{E}(Y | X, B)|$ for each $X_i \in X$. The non-linear regression residuals are equivalent to the error terms assuming an additive noise model.
2. Linear Non-Gaussian Acyclic Model (LiNGAM) (Peters et al., 2014; Strobl and Lasko, 2022): same as above but performs linear instead of non-linear regression.
3. CausalCell (Wen et al., 2023): selects the top 50 genes with maximal statistical dependence to Y , and then runs the Peter-Clark (PC) algorithm (Spirtes et al., 2000) using a non-parametric conditional independence test to identify a causal graph among the top 50 genes. The algorithm does not perform root causal inference, so we use ANM as above but condition on the estimated parent sets for the top 50 genes and the ancestors inferred from 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 $X_i \in X$.
5. Multivariate regression residuals (Multi Reg): similar to above but instead computes the absolute residuals after regressing Y on $(X \setminus X_i) \cup B$.

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 (Friedman, 1991). We also standardized all variables before running the regressions to prevent gaming of the marginal variances in causal discovery (Reisach et al., 2021; Ng et al., 2024). 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 undifferentiated blast cells from the K562 cell line (Replogle et al., 2022 [\[1\]](#)). We used the genomewide dataset version for the latter. We downloaded the datasets from the scPerturb database on Zenodo (Green et al., 2022 [\[2\]](#)) 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) (Ratnapriya et al., 2019 [\[3\]](#); Kim et al., 2021 [\[4\]](#)). 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 Perturbseq dataset, yielding a final count of 513 bulk RNA-seq samples and 247,914 Perturb-seq samples across 2,077 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 6,882 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 with large causal effect in diseases that progress over time.
Second, few root causal genes should drive pathogenesis because the effects of a few error terms distribute over many downstream genes. We verify the sparsity of root causal genes 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 $\tilde{X}_i \in \tilde{X}$ where $\Omega_i = |\mathbb{E}(Y | X_i, B) - \mathbb{E}(Y | B)|$ and ω_{ij} its value for patient j .
Despite the sparsity 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 the vast majority of 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 nonlinear regression method, so we instead estimated the expectations using kernel ridge regression equipped with a radial basis function kernel (Shawe-Taylor and Cristianini, 2004 [↗](#)). We then computed the D-RCS across all patients for each variable in X . We ran pathway enrichment analysis using the fast gene set enrichment analysis (FGSEA) algorithm (Sergushichev, 2016 [↗](#)) with one hundred thousand simple permutations using the D-RCS scores and pathway information from the Reactome database (version 1.86.0) (Fabregat et al., 2017 [↗](#)). We likewise performed drug set enrichment analysis with the Drug Signature database (version 1.0) (Yoo et al., 2015 [↗](#)). We repeated the above procedures for the D-RCS of all clusters identified by hierarchical clustering via Ward's method (Ward Jr, 1963 [↗](#)).

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* [↗](#).

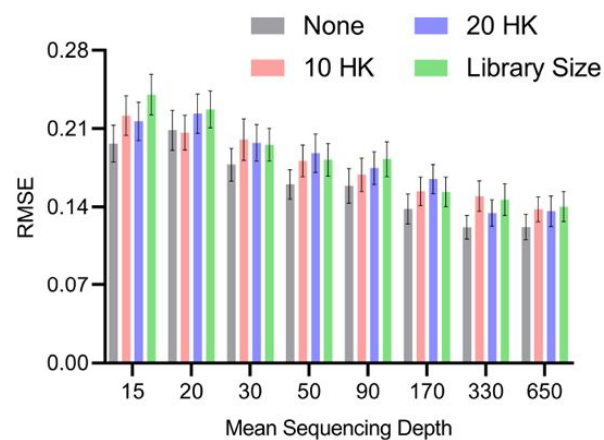
Supplementary Materials

Additional Synthetic Data Results

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 no normalization exceeds or matches the accuracy of other strategies. We therefore do not normalize by sequencing depth in subsequent analyses.



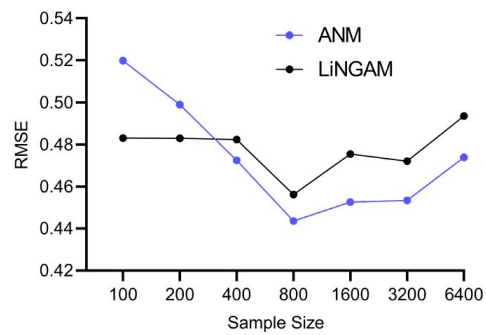
Supplementary Figure 1.

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.

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 2** [↗](#). The accuracies of ANM and LiNGAM did not improve beyond an RMSE of 0.44 to the ground truth error term values 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.



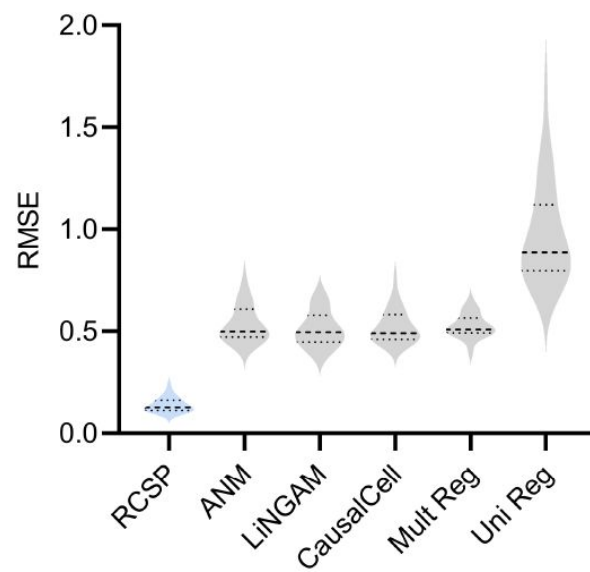
Supplementary Figure 2.

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.

Cyclic Causal Graphs

We also evaluated the algorithms on directed graphs with cycles. We generated a linear SEM over $p + 1 = 1000$ variables in $\tilde{X} \cup Y$. We sampled the coefficient matrix β from a Bernoulli($1/(p - 1)$) distribution but did not restrict the non-zero coefficients to the upper triangular portion of the matrix. We then proceeded to permute the variable ordering and weight each entry as in the Methods for the DAG. We repeated this procedure 30 times and report the results in **Supplementary Figure 3** [↗](#).

RCSP again outperformed all other algorithms even in the cyclic case. The results suggest that conditioning on the surrogate ancestors also estimates the RCS well even in the cyclic case. However, we caution that an error term E_i can affect the ancestors of $\tilde{X}_i$ when cycles exist. As a result, the RCS may not isolate the causal effect of the error term and thus not truly coincide with the notion of a root causal effect in cyclic causal graphs.



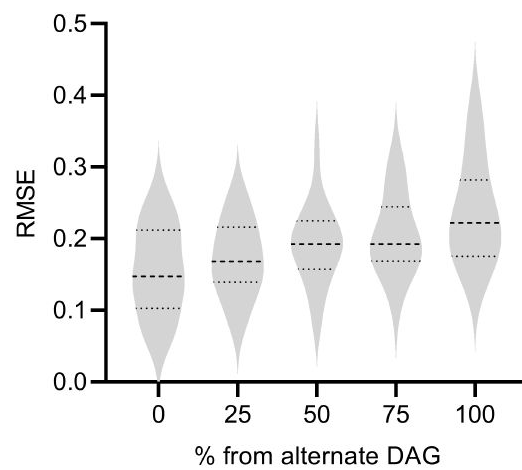
Supplementary Figure 3.

RCSP achieved the lowest RMSE in cyclic graphs as well. However, error terms can influence ancestors in the cyclic case, so the interpretation of the RCS remains unclear when cycles exist.

DAG Incongruence

We next assessed the performance of RCSP when the DAG underlying the Perturb-seq data differs from the DAG underlying the bulk RNA-seq data. We considered a mixture of two random DAGs in bulk RNA-seq, where one of the DAGs coincided with the Perturb-seq DAG and second alternate DAG did not. We instantiated and simulated samples from each DAG as per the previous subsection. We generated 0%, 25%, 50%, 75%, and 100% of the bulk RNA-seq samples from the alternate DAG, and the rest from the Perturb-seq DAG. We ideally would like to see the performance of RCSP degrade gracefully, as opposed to abruptly, as the percent of samples derived from the alternate DAG increases.

We summarize results in **Supplementary Figure 4** [↗](#). As expected, RCSP performed the best when we drew all samples from the same underlying DAG for Perturb-seq and bulk RNA-seq. However, the performance of RCSP also degraded slowly as the percent of samples increased from the alternate DAG. We conclude that RCSP can accommodate some differences between the underlying DAGs in Perturb-seq and bulk RNA-seq with only a mild degradation in performance.

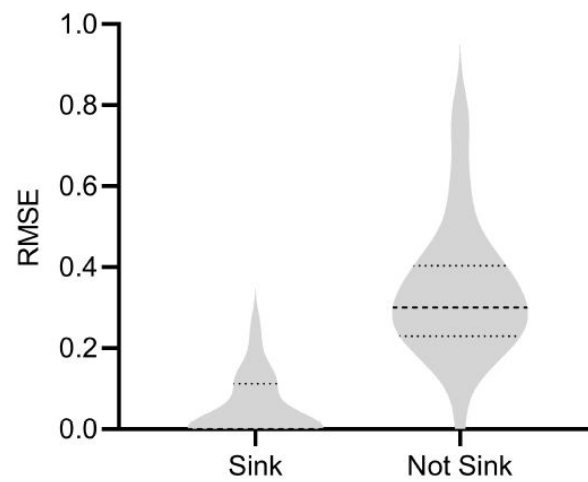


Supplementary Figure 4.

The performance of RCSP degrades gracefully as the percent of samples from the alternate DAG increases.

Non-Sink Target

We finally considered the scenario where Y is a non-sink (or non-terminal) vertex. If Y is a parent of a gene expression level, then we cannot properly condition on the parents because modern Perturb-seq datasets usually do not intervene on Y or measure Y . We therefore empirically investigated the degradation in performance resulting from a non-sink target Y , in particular for gene expression levels where Y is a parent. We again simulated 200 samples from bulk RNA-seq and each condition of Perturb-seq with a DAG over 1000 vertices, an expected neighborhood size of 2 and a non-sink target Y . We then removed the outgoing edges from Y and resampled the DAG with a sink target. We compare the results of RCSP for both DAGs in gene expression levels where Y is a parent. We plot the results in **Supplementary Figure 5** [↗](#). As expected, we observe a degradation in performance when Y is not terminal, where the mean RMSE increased from 0.045 to 0.342. We conclude that RCSP is sensitive to violations of the sink target assumption.



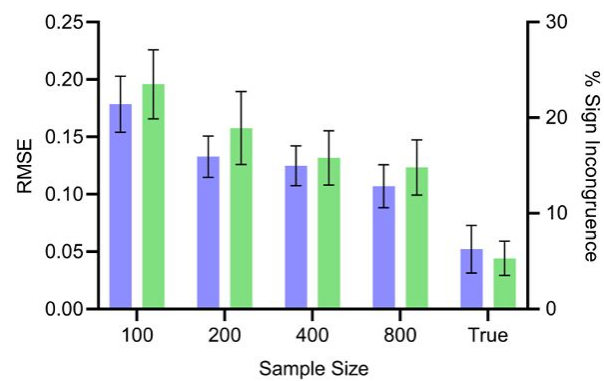
Supplementary Figure 5.

Results with a sink or non-sink target Y . RCSP estimated the RCS scores less accurately with a non-sink target indicating that the algorithm is sensitive to violations of the sink target assumption.

Root Causal Effect versus Conditional Root Causal Effect

We compared the expected and unconditional root causal effects $\Omega_i \triangleq \mathbb{E}(Y \mid E_i) - \mathbb{E}(Y)$ to the expected and conditional root causal effects or, equivalently, the signed RCS scores Γ . These root causal effects and conditional root causal effects are not the same, but they are similar. We empirically investigated the differences between the estimated values of Γ and the true values of Ω using the RMSE and also the percent of samples with incongruent signs; Γ and Ω 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 Γ against true Ω by estimating the two to negligible error using 20,000 samples of $\tilde{\mathbf{X}}$. We repeated each experiment 100 times and thus generated a total of $100 \times 5 = 500$ datasets.

We summarize the results in **Supplementary Figure 6** [↗](#). The estimated Γ values approached the true Ω values with increasing sample sizes. The true Γ values did not converge exactly to the true Ω 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 $\tilde{\mathbf{X}}$ to 50,000 did not change performance, confirming that we reached the floor. We conclude that the empirical results replicate the theoretical results because Γ and Ω 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 6.

Mean RMSE (blue, left) and percent sign incongruence (green, right) of the expected root causal effects and signed RCS values, respectively. 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%.

Additional Results for Age-related Macular Degeneration

Algorithm Comparisons

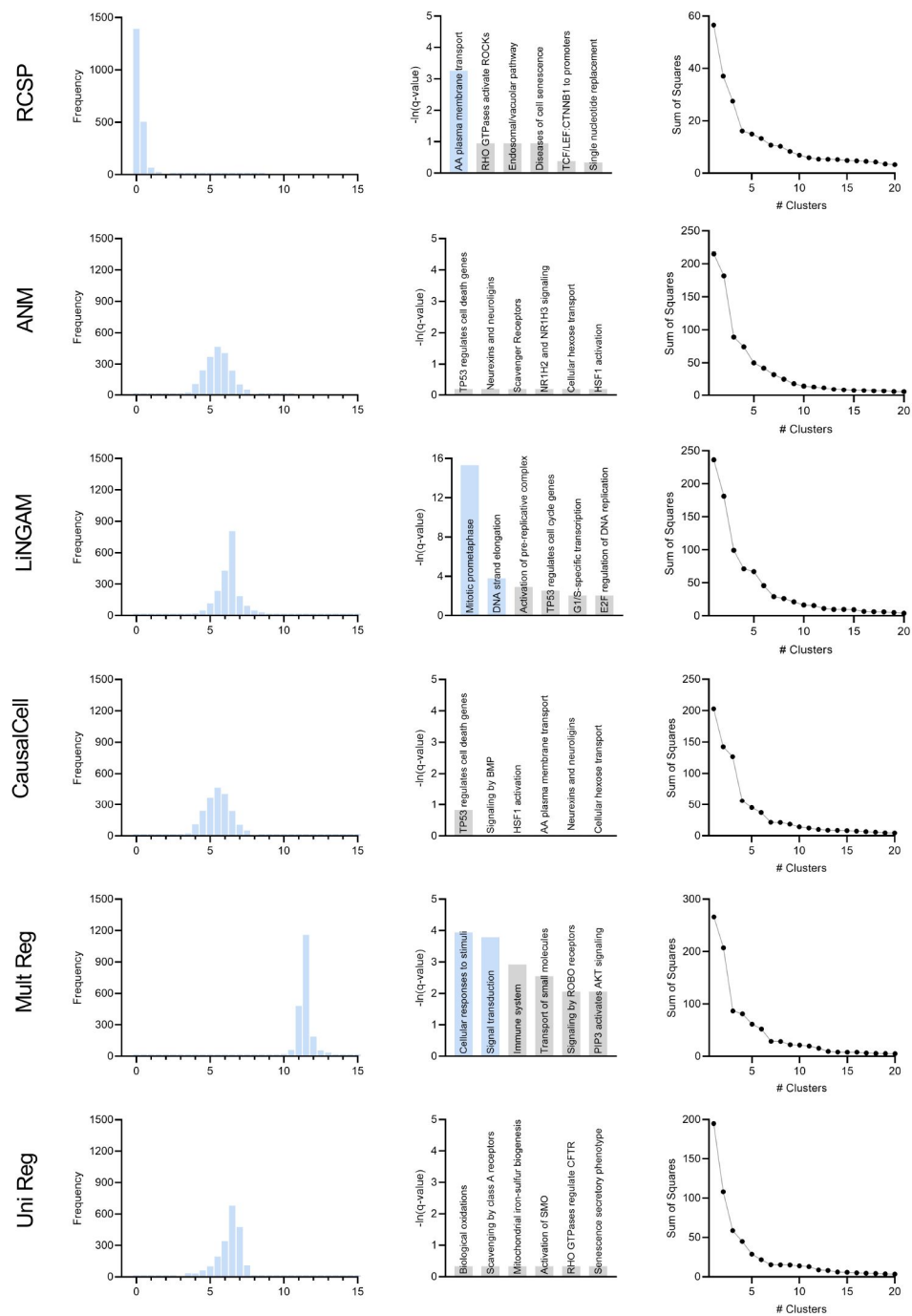
We say that an algorithm performs well in real data if it simultaneously (1) identifies a sparse set of root causal genes, (2) recovers known pathogenic pathways with high specificity measured by the sparsity of leading edge genes, and (3) clusters patients into clear subgroups.

We compared the algorithms with the AMD data. We summarize the results in **Supplementary Figure 7** [↗](#) plotted on the next page. The figure contains 6 rows and 3 columns. Similar to the D-RCS, we can compute the standard deviation of the output of each algorithm from zero for each gene. The first column in **Supplementary Figure 7** [↗](#) denotes the histograms of these standard deviations across the genes. We standardized the outputs to have mean zero and unit variance. We then added the minimum value so that all histograms begin at zero; note that the bars at zero are not visible for many algorithms, since only a few genes attained standard deviations near the minimum. If an algorithm accurately identifies root causal genes, then it should only identify a few genes with large conditional root causal effects under the omnigenic root causal model. The RCSP algorithm had a histogram with large probability mass centered around zero with a long tail to the right. The standard deviations of the outputs of the other algorithms attained large values for nearly all genes. 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 7** [↗](#). 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 7** [↗](#). The RCSP sum of squares plot revealed a sharp elbow at four 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 a small set of root causal genes, identified pathogenic pathways with maximal specificity and discovered distinguishable patient subgroups. We therefore conclude that RCSP outperformed all other algorithms in the AMD dataset.

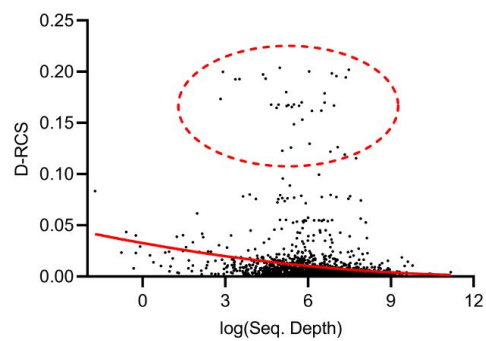


Supplementary Figure 7.

Comparison of the algorithms in age-related macular degeneration.

Effect of Sequencing Depth

Theorem 1 states that RCS scores may exhibit bias with insufficient sequencing depth. The genes with large D-RCS scores may therefore simply have low sequencing depths. To test this hypothesis, we plotted sequencing depth against D-RCS scores. Consistent with Theorem 1, we observed a small negative correlation between D-RCS and sequencing depth ($\rho = -0.16$, $p=2.04E-13$), and D-RCS scores exhibited greater variability at the lowest sequencing depths (**Supplementary Figure 8** [↗](#)). However, genes with the largest D-RCS scores had mean sequencing depths interspersed between 20 and 3000. We conclude that genes with the largest D-RCS scores had a variety of sequencing depths ranging from low to high.



Supplementary Figure 8.

Mean sequencing depth of each gene plotted against their D-RCS scores in AMD. Genes with the largest D-RCS scores (red ellipse) had a variety of sequencing depths.

Biological Results

We provide the full pathway enrichment analysis results in **Supplementary Table 1** [corresponding to Figure 3 \(c\)](#). We summarize pathway enrichment analysis of the black cluster of **Figure 3 (g)** [in Figure 3 \(j\)](#). However, analyses of the blue, green and pink clusters did not yield significant pathways even at a liberal FDR threshold of 10%.

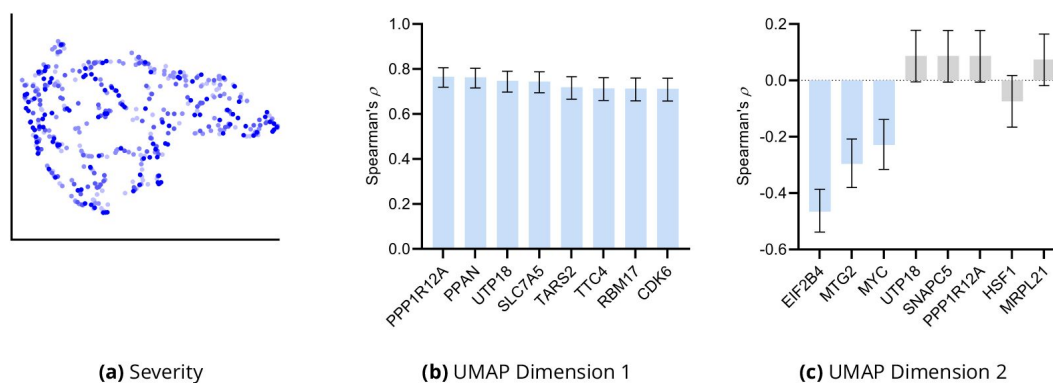
We examined whether the clusters of **Figure 3 (g)** [differentiate dry and wet macular degeneration](#). Wet macular degeneration is associated with the highest Minnesota Grading System (MGS) score of 4 (Olsen and Feng, 2004 [Figure 9 \(a\)](#)). We plotted the UMAP embedding against MGS (**Supplementary Figure 9 (a)** [Figure 3 \(a\)](#)). 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 3 (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.

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
PI3P 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.

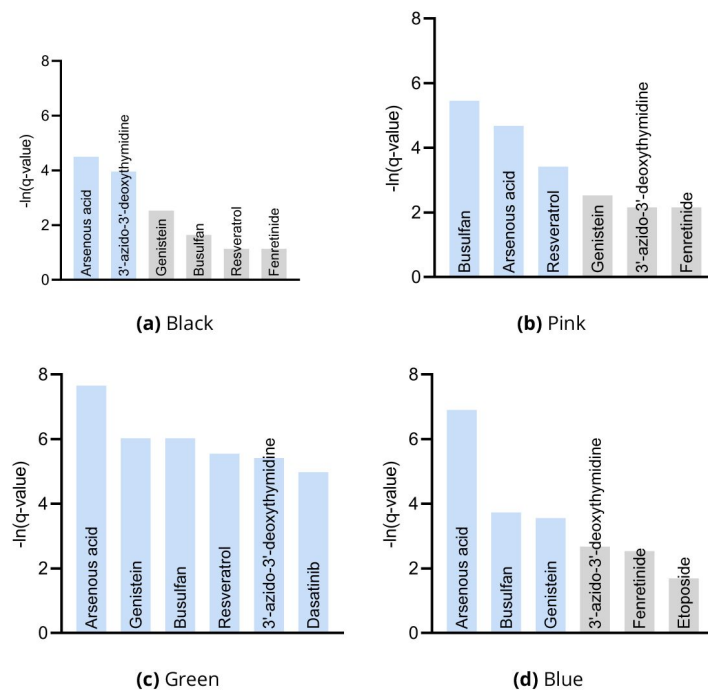
We correlated the two UMAP dimensions with the top 30 genes ranked by their RCS scores. We plot genes with the highest correlation to the first and second UMAP dimensions in **Supplementary Figures 9 (b)** [↗](#) and **9 (c)** [↗](#), respectively. Many genes correlated with the first dimension, but only three genes correlated with the second at an FDR threshold of 5%.



Supplementary Figure 9.

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 3 (g)** [↗](#). We summarize the results in **Supplementary Figure 10** [↗](#). 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 10.

Drug enrichment analysis results by cluster in **Figure 3** (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

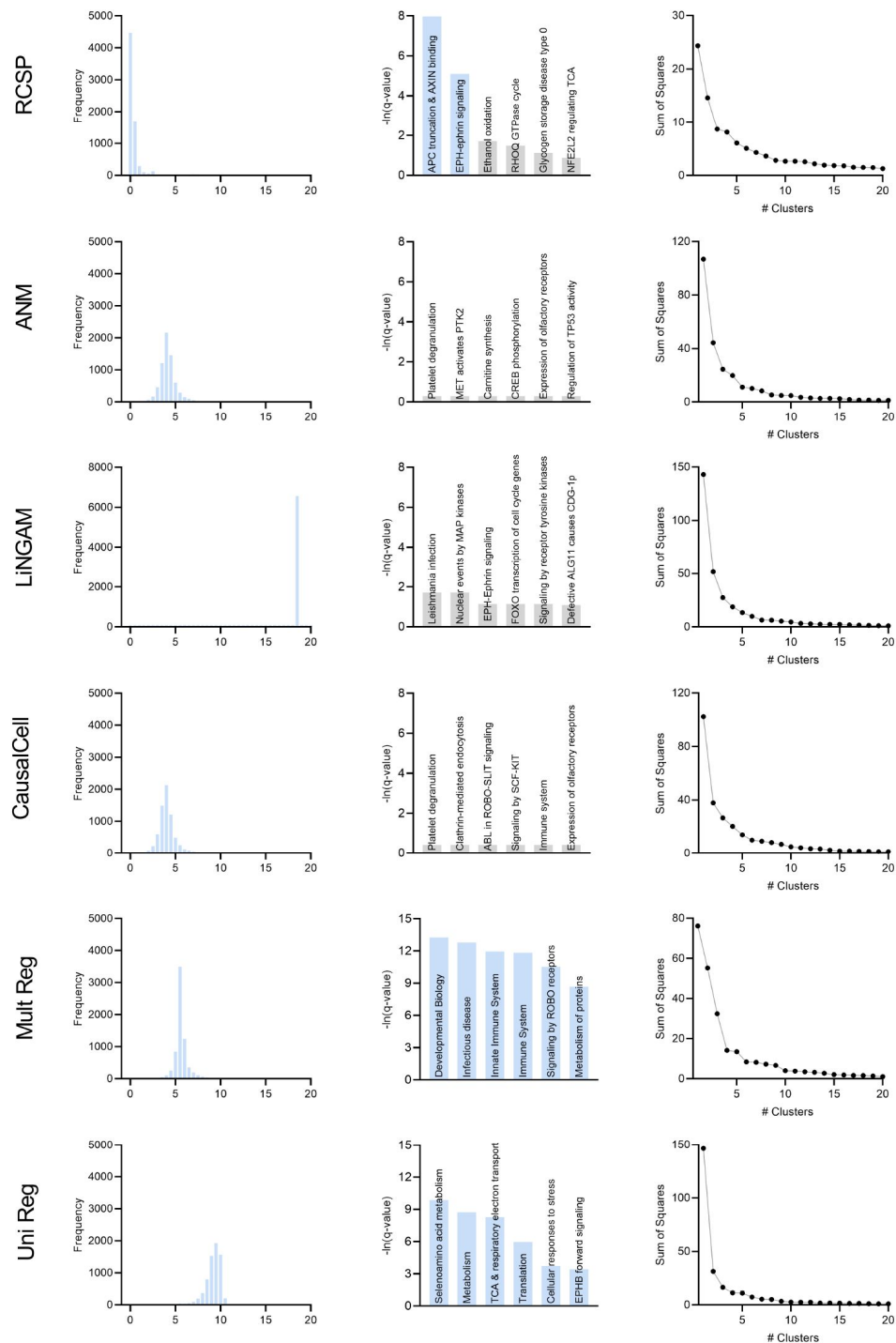
Algorithm Comparisons

We compared the algorithms using the MS data with the same criteria used for the AMD dataset. We summarize the results in **Supplementary Figure 11** [↗](#) 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 11** [↗](#). 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 11** [↗](#). 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.

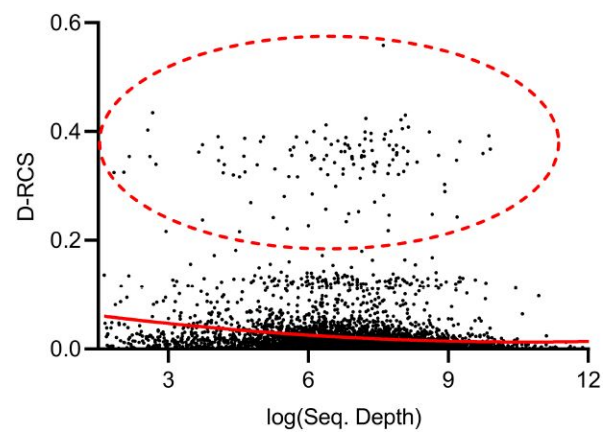


Supplementary Figure 11.

Comparison of the algorithms in multiple sclerosis.

Effect of Sequencing Depth

We plot sequencing depth against the D-RCS scores of each gene similar to the AMD dataset. We again observed a small negative correlation ($\rho = -0.136$, $p < 2.2E-16$), indicating that genes with low sequencing depths had slightly higher D-RCS scores on average (**Supplementary Figure 12** [↗](#)). However, genes with the largest D-RCS scores again had a variety of sequencing depths. We conclude that sequencing depth has minimal correlation with the largest D-RCS scores.



Supplementary Figure 12.

Mean sequencing depth of each gene plotted against their D-RCS scores in MS. Genes with the largest D-RCS scores (red ellipse) again had a variety of sequencing depths.

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 4 (f)** [↗](#) consistently implicated EPH-ephrin signaling among the top two pathways. However, each cluster also involved one separate additional pathway (**Supplementary Figure 13** [↗](#)). 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 (Orlan et al., 2021 [↗](#)). Finally, CTLA4 regulates T-cell homeostasis and inhibits autoimmunity for the pink cluster (Basile et al., 2022 [↗](#)). The D-RCS scores of each cluster thus implicate different mechanisms of T cell pathology.

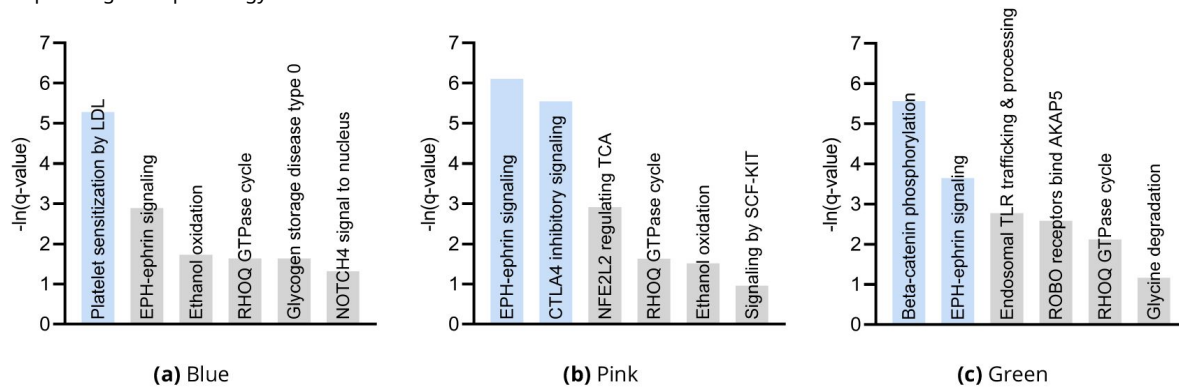
Supplementary Table 2.

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.

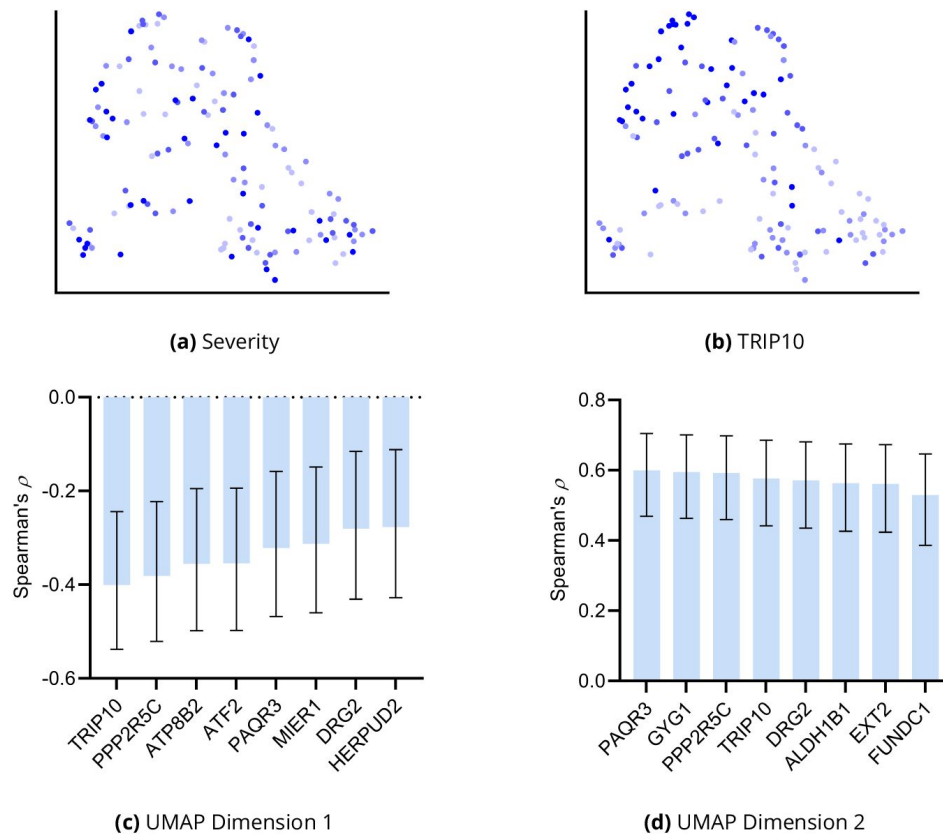
Pathway	p-value	q-value	Effect Size	Leading Edge
APC truncation mutants have impaired AXIN binding	1.91e-06	3.45e-4	0.960	5525,5527
EPH-ephrin signaling	4.23e-05	6.12e-3	0.826	8874,102,8976
Ethanol oxidation	2.02e-03	0.182	0.967	219,128
RHOQ GTPase cycle	2.72e-03	0.226	0.793	9322,8874,10395
Glycogen 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 Figure 13.

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 14 (a)** [↗](#)). The top genes in **Figure 4** [↗](#) (d) such as MNT and CERCAM also did not correlate. However, lower ranked genes such as TRIP10 did (**Supplementary Figure 14 (b)** [↗](#)). An expanded correlation analysis with the top 30 genes revealed significant correlations across a variety of lower ranked genes (**Supplementary Figures 14 (c)** [↗](#) and **14 (d)** [↗](#)). We conclude that the distribution of lower ranked genes govern the structure of the UMAP embedding in **Figure 4 (f)** [↗](#).



Supplementary Figure 14.

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 $N \geq n_0$:

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

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

Proof. We can write the following sequence:

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

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 $\alpha > 0$ and invoke the Markov inequality:

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

The conclusion follows because we chose α arbitrarily.

Proposition 1.

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

Proof. If $E_i \not\perp\!\!\!\perp Y$ or $E_i \not\perp\!\!\!\perp Y|\text{Pa}(\tilde{X}_i)$ (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 $\mathbb{P}(Y|E_i, \text{Pa}(\tilde{X}_i)) = \mathbb{P}(Y|\tilde{X}_i, \text{Pa}(\tilde{X}_i))$ under Equation (3).

Proof. We can write:

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

The second equality follows because $\tilde{X}_i$ is a constant given E_i and $\text{Pa}(\tilde{X}_i)$. 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 $N \rightarrow \infty$.

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, $SD(\tilde{X}_i)$ contains only and all the surrogate descendants of $\tilde{X}_i$ for each $\tilde{X}_i \in \tilde{\mathbf{X}}$. This in turn implies that $SA(\tilde{X}_i)$ in Line 5 of Algorithm 1 contains only and all the surrogate ancestors of. Hence, RCSP now has $\tilde{X}_i$ access to the correct set $SA(\tilde{X}_i)$ as well as B for each $\tilde{X}_i \in \tilde{\mathbf{X}}$. We finally invoke Theorem 1 to conclude that RCSP recovers Φ almost surely as $N \rightarrow \infty$.

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.

References

- Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, et al. 2016) **A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response** *Cell* **167**:1867–1882
- Andhavarapu S, Mubariz F, Arvas M, Bever Jr C, Makar TK 2019) **Interplay between ER stress and autophagy: a possible mechanism in multiple sclerosis pathology** *Experimental and Molecular Pathology* **108**:183–190
- Barouch FC, Miller JW 2007) **The role of inflammation and infection in age-related macular degeneration** *International ophthalmology clinics* **47**:185–197
- Basile MS, Bramanti P, Mazzon E. 2022) **The role of cytotoxic T-lymphocyte antigen 4 in the pathogenesis of multiple sclerosis** *Genes* **13**:1319
- Beaumat F, O’prey J, Barthet VJ, Zunino B, Parvy JP, Bachmann AM, O’prey M, Kania E, Gonzalez PS, Macintosh R, et al. 2019) **mTORC1 activation requires DRAM-1 by facilitating lysosomal amino acid efflux** *Molecular Cell* **76**:163–176
- Bongers S, Forré P, Peters J, Mooij JM 2021) **Foundations of structural causal models with cycles and latent variables** *The Annals of Statistics* **49**:2885–2915
- Boyle EA, Li YI, Pritchard JK 2017) **An expanded view of complex traits: from polygenic to omnigenic** *Cell* **169**:1177–1186
- Burster T, Beck A, Poeschel S, Øren A, Baechle D, Reich M, Roetzschke O, Falk K, Boehm BO, Youssef S, et al. 2007) **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
- Buschur KL, Chikina M, Benos PV 2020) **Causal network perturbations for instance-specific analysis of single cell and disease samples** *Bioinformatics* **36**:2515–2521
- Cano-Gamez E, Trynka G. 2020) **From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases** *Frontiers in Genetics* **11**:424
- Choudhary S, Satija R. 2022) **Comparison and evaluation of statistical error models for scRNA-seq** *Genome Biology* **23**:27
- Colombo D, Maathuis MH, et al. 2014) **Order-independent constraint-based causal structure learning** *Journal of Machine Learning Research* **15**:3741–3782
- Costa-Silva J, Domingues D, Lopes FM 2017) **RNA-Seq differential expression analysis: An extended review and a software tool** *PloS One* **12**:e0190152
- Dalvin LA, Olsen TW, Bakri SJ, McCullough K, Tefferi A, Al-Kali A. 2022) **Busulfan treatment for myeloproliferative disease may reduce injection burden in vascular endothelial growth factor-driven retinopathy** *American Journal of Ophthalmology Case Reports* **26**:101554

Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. 2017) **Pooled CRISPR screening with single-cell transcriptome readout** *Nature methods* **14**:297–301

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. 2016) **Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens** *Cell* **167**:1853–1866

Ellington CN, Lengerich BJ, Watkins TB, Yang J, Xiao H, Kellis M, Xing EP 2023) **Contextualized Networks Reveal Heterogeneous Transcriptomic Regulation in Tumors at Sample-Specific Resolution** *Neural Information and Processing Systems Workshop on Generative AI and Biology*

Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, D'Eustachio P, Stein L, Hermjakob H. 2017) **Reactome pathway analysis: a high-performance in-memory approach** *BMC Bioinformatics* **18**:1–9

Fletcher JM, Lalor S, Sweeney C, Tubridy N, Mills K. 2010) **T cells in multiple sclerosis and experimental autoimmune encephalomyelitis** *Clinical & Experimental Immunology* **162**:1–11

Friedman JH 1991) **Multivariate adaptive regression splines** *The Annals of Statistics* **19**:1–67

Friedman N, Linial M, Nachman I, Pe'er D. 2000) **Using Bayesian networks to analyze expression data** *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology* :127–135

Gnanaprakasam JR, Wang R. 2017) **MYC in regulating immunity: metabolism and beyond** *Genes* **8**:88

Go YM, Zhang J, Fernandes J, Litwin C, Chen R, Wensel TG, Jones DP, Cai J, Chen Y. 2020) **MTOR-initiated metabolic switch and degeneration in the retinal pigment epithelium** *FASEB Journal* **34**:12502

Golan M, Krivitsky A, Mausner-Fainberg K, Benhamou M, Vigiser I, Regev K, Kolb H, Karni A. 2021) **Increased expression of ephrins on immune cells of patients with relapsing remitting multiple sclerosis affects oligodendrocyte differentiation** *International Journal of Molecular Sciences* **22**:2182

Green TD, Peidli S, Shen C, Gross T, Min J, Garda S, Taylor-King JP, Marks DS, Luna A, Blüthgen N, et al. 2022) **scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data** *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*

Grün D, Kester L, Van Oudenaarden A. 2014) **Validation of noise models for single-cell transcriptomics** *Nature Methods* **11**:637–640

Hadziahmetovic M, Malek G. 2021) **Age-related macular degeneration revisited: From pathology and cellular stress to potential therapies** *Frontiers in Cell and Developmental Biology* **8**:612812

Haves-Zburof D, Paperna T, Gour-Lavie A, Mandel I, Glass-Marmor L, Miller A. 2011) **Cathepsins and their endogenous inhibitors cystatins: expression and modulation in multiple sclerosis** *Journal of Cellular and Molecular Medicine* **15**:2421–2429

Kamalden T, Ji D, Fawcett R, Osborne N. 2011) **Genistein blunts the negative effect of ischaemia to the retina caused by an elevation of intraocular pressure** *Ophthalmic*

Kim K, Pröbstel AK, Baumann R, Dyckow J, Landefeld J, Kogl E, Madireddy L, Loudermilk R, Eggers EL, Singh S, et al. 2021) **Cell type-specific transcriptomics identifies neddylation as a novel therapeutic target in multiple sclerosis** *Brain* **144**:450–461

Kinoshita S, Noda K, Tagawa Y, Inafuku S, Dong Y, Fukuhara J, Dong Z, Ando R, Kanda A, Ishida S. 2014) **Genistein attenuates choroidal neovascularization** *The Journal of Nutritional Biochemistry* **25**:1177–1182

Kokame K, Agarwala KL, Kato H, Miyata T. 2000) **Herp, a new ubiquitin-like membrane protein induced by endoplasmic reticulum stress** *Journal of Biological Chemistry* **275**:32846–32853

Lengfeld JE, Lutz SE, Smith JR, Diaconu C, Scott C, Kofman SB, Choi C, Walsh CM, Raine CS, Agalliu I, et al. 2017) **Endothelial Wnt/ β -catenin signaling reduces immune cell infiltration in multiple sclerosis** *Proceedings of the National Academy of Sciences* **114**:E1168–E1177

Luo H, Broux B, Wang X, Hu Y, Ghannam S, Jin W, Larochelle C, Prat A, Wu J. 2016) **EphrinB1 and EphrinB2 regulate T cell chemotaxis and migration in experimental autoimmune encephalomyelitis and multiple sclerosis** *Neurobiology of Disease* **91**:292–306

Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H, et al. 2020) **A compendium of mutational cancer driver genes** *Nature Reviews Cancer* **20**:555–572

Nachef M, Ali AK, Almutairi SM, Lee SH 2021) **Targeting SLC1A5 and SLC3A2/SLC7A5 as a potential strategy to strengthen anti-tumor immunity in the tumor microenvironment** *Frontiers in immunology* **12**:624324

Nagral A. 2014) **Gaucher disease** *Journal of Clinical and Experimental Hepatology* **4**:37–50

Narendran S, Pereira F, Yerramothu P, Apicella I, Wang Sb, Varshney A, Baker KL, Marion KM, Ambati M, Ambati VL, et al. 2020) **A clinical metabolite of azidothymidine inhibits experimental choroidal neovascularization and retinal pigmented epithelium degeneration** *Investigative ophthalmology & visual science* **61**:4–4

Ng I, Huang B, Zhang K. 2024) **Structure learning with continuous optimization: A sober look and beyond** *PMLR* :71–105

Nicklin P, Bergman P, Zhang B, Triantafellow E, Wang H, Nyfeler B, Yang H, Hild M, Kung C, Wilson C, et al. 2009) **Bidirectional transport of amino acids regulates mTOR and autophagy** *Cell* **136**:521–534

Olsen TW, Feng X. 2004) **The Minnesota Grading System of eye bank eyes for age-related macular degeneration** *Investigative Ophthalmology and Visual Science* **45**:4484–4490

Orian JM, D’Souza CS, Kocovski P, Krippner G, Hale MW, Wang X, Peter K. 2021) **Platelets in multiple sclerosis: early and central mediators of inflammation and neurodegeneration and attractive targets for molecular imaging and site-directed therapy** *Frontiers in Immunology* **12**:620963

Papoulis A. 1984) **Probability, Random Variables and Stochastic Processes** McGraw-Hill

- Pearl J. 2009) **Causality** Cambridge University Press
- Peters J, Mooij JM, Janzing D, Schölkopf B. 2014) **Causal discovery with continuous additive noise models** *Journal of Machine Learning Research*
- Ratnapriya R, Sosina OA, Starostik MR, Kwicklis M, Kapphahn RJ, Fritsche LG, Walton A, Arvanitis M, Gieser L, Pietraszkiewicz A, et al. 2019) **Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration** *Nature Genetics* **51**:606–610
- Reisach A, Seiler C, Weichwald S. 2021) **Beware of the simulated DAG! causal discovery benchmarks may be easy to game** *Advances in Neural Information Processing Systems* **34**:27772–27784
- Replogle JM, Saunders RA, Pogson AN, Hussmann JA, Lenail A, Guna A, Mascibroda L, Wagner EJ, Adelman K, Lithwick-Yanai G, et al. 2022) **Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq** *Cell* **185**:2559–2575
- Sarkar A, Stephens M. 2021) **Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis** *Nature Genetics* **53**:770–777
- Sergushichev A. 2016) **An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation** *BioRxiv* **60012**:1–9
- Shawe-Taylor J, Cristianini N. 2004) **Kernel Methods for Pattern Analysis** Cambridge University Press
- Shi Y, Qu J, Zhang D, Zhao P, Zhang Q, Tam POS, Sun L, Zuo X, Zhou X, Xiao X, et al. 2011) **Genetic variants at 13q12. 12 are associated with high myopia in the Han Chinese population** *The American Journal of Human Genetics* **88**:805–813
- Sobel RA 2005) **Ephrin A receptors and ligands in lesions and normal-appearing white matter in multiple sclerosis** *Brain Pathology* **15**:35–45
- Spink KE, Polakis P, Weis WI 2000) **Structural basis of the Axin–adenomatous polyposis coli interaction** *The EMBO journal* **19**:2270–2279
- Spirtes P, Glymour C, Scheines R. 2000) **Causation, Prediction, and Search** MIT press
- Spirtes P. 1995) **Directed cyclic graphical representations of feedback models** *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* :491–498
- Starzyk RM, Rosenow C, Frye J, Leismann M, Rodzinski E, Putney S, Tuomanen EI 2000) **Cerebral cell adhesion molecule: a novel leukocyte adhesion determinant on blood-brain barrier capillary endothelium** *The Journal of Infectious Diseases* **181**:181–187
- Strobl EV 2023) **Causal discovery with a mixture of DAGs** *Machine Learning* **112**:4201–4225
- Strobl EV 2024) **Counterfactual Formulation of Patient-Specific Root Causes of Disease** *Journal of Biomedical Informatics*
- Strobl EV, Lasko TA 2022) **Identifying patient-specific root causes of disease** *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* :1–10

Strobl EV, Lasko TA (2023) **Identifying patient-specific root causes with the heteroscedastic noise model** *Journal of Computational Science* **72**:102099

Strobl EV, Lasko TA (2023) **Root Causal Inference from Single Cell RNA Sequencing with the Negative Binomial** *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics BCB '23* New York, NY, USA: Association for Computing Machinery

Strobl EV, Lasko TA (2023) **Sample-specific root causal inference with latent variables** *PMLR* :895–915

Strobl EV, Lasko TA, Gamazon ER (2024) **Mitigating pathogenesis for target discovery and disease subtyping** *Computers in Biology and Medicine* **171**:108122

Su Y, Wang F, Hu Q, Qu Y, Han Y. (2020) **Arsenic trioxide inhibits proliferation of retinal pigment epithelium by down-regulating expression of extracellular matrix and p27** *International Journal of Clinical and Experimental Pathology* **13**:172

Turi Z, Senkyrikova M, Mistrik M, Bartek J, Moudry P. (2018) **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

Wang L, Trasanidis N, Wu T, Dong G, Hu M, Bauer DE, Pinello L. (2023) **Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics** *Nature Methods* **20**:1368–1378

Ward Jr JH (1963) **Hierarchical grouping to optimize an objective function** *Journal of the American Statistical Association* :236–244

Wen Y, Huang J, Guo S, Elyahu Y, Monsonogo A, Zhang H, Ding Y, Zhu H. (2023) **Applying causal discovery to single-cell analyses using CausalCell** *Elife* **12**:e81464

Yoo M, Shin J, Kim J, Ryall KA, Lee K, Lee S, Jeon M, Kang J, Tan AC (2015) **DSigDB: drug signatures database for gene set analysis** *Bioinformatics* **31**:3069–3071

Author information

Eric V Strobl

University of Pittsburgh, Pittsburgh, United States

For correspondence: eric.strobl@pitt.edu

Eric R Gamazon

Vanderbilt University Medical Center, Nashville, United States

Editors

Reviewing Editor

David Knowles

Department of Computer Science and Department of Systems Biology, New York, United States of America

Senior Editor

Alan Moses

University of Toronto, Toronto, Canada

Reviewer #1 (Public review):

Summary:

This manuscript seeks to estimate the causal effect of genes on disease. To do so, they introduce a novel algorithm, termed the Root Causal Strength using Perturbations (RCSP) algorithm. RCSP uses perturb-seq to first estimate the gene regulatory network structure among genes, and then uses bulk RNA-seq with phenotype data on the samples to estimate causal effects of genes on the phenotype conditional on the learned network structure. The authors assess the performance of RCSP in comparison to other methods via simulation. Next, they apply RCSP to two real human datasets: 513 individuals age-related macular degeneration and 137 individuals with multiple sclerosis.

Strengths:

The authors tackle an important and ambitious problem - the identification of causal contributors to disease in the context of a causal inference framework. As the authors point out, observational RNA-seq data is insufficient for this kind of causal discovery, since it is very challenging to recover the true underlying graph from observational data; interventional data are needed. However, little perturb-seq data has been generated with annotated phenotype data, and much bulk RNA-seq data has already been generated, so it is useful to propose an algorithm to integrate the two as the authors have done.

The authors also offer substantial theoretical exposition for their work, bringing to bear both the literature on causal discovery as well as literature on the genetic architecture of complex traits. They also benchmark RCSP under multiple challenging simulation settings, including an analysis of RCSP when the underlying graph is not a DAG.

Weaknesses:

The notion of a "root" causal gene - which the authors define based on a graph theoretic notion of topologically sorting graphs - requires a graph that is directed and acyclic. It is the latter that constitutes an important weakness here - it simply is a large simplification of human biology to draw out a DAG including hundreds of genes and a phenotype Y and to claim that the true graph contains no cycles. For example - consider the authors' analysis of T cell infiltration in multiple sclerosis (MS). CD4⁺ effector T cells have the interesting property that they are stimulated by IL2 as a growth factor; yet IL2 also stimulates the activation of (suppressive) regulatory T cells. What does it mean to analyze CD4⁺ regulation in disease with a graph that does not consider IL2 (or other cytokine) mediated feedback loops/cycles? To the authors' credit, in the supplementary materials they do consider a simulated example with a cyclic underlying causal graph, finding that RCSP performed well comparison to an implementation of the additive noise model (ANM), LiNGAM, CausalCell, and two simpler approaches based on linear regression.

I also encourage the authors to consider more carefully when graph structure learned from perturb-seq can be ported over to bulk RNA-seq. Consider again the MS CD4+ example - the authors first start with a large perturb-seq experiment (Replogle et al., 2022) performed in K562 cells. To what extent are K562 cells, which are derived from a leukemia cell line, suitable for learning the regulatory structure of CD4+ cells from individuals with an MS diagnosis? Presumably this structure is not exactly correct - to what extent is the RCSP algorithm sensitive to false edges in this graph? The authors perform an analysis of this scenario in Supplementary Figure 4, which shows that RCSP is robust to some degree of departure from the underlying true structure. And although challenging - it would be ideal for the RCSP to model or reflect the challenges in correctly identifying the regulatory structure.

It should also be noted that in most perturb-seq experiments, the entire genome is not perturbed, and frequently important TFs (that presumably are very far "upstream" and thus candidate "root" causal genes) are not expressed highly enough to be detected with scRNA-seq. In that context - perhaps slightly modifying the language regarding RCSP's capabilities might be helpful for the manuscript - perhaps it would be better to describe it has an algorithm for causal discovery among a set of genes that were perturbed and measured, rather than a truly complete search for causal factors. Perhaps more broadly - it would also benefit the manuscript to devote slightly more text to describing the kinds of scenarios where RCSP (and similar ideas) would be most appropriately applied - perhaps a well-powered, phenotype annotated perturb-seq dataset performed in a disease relevant primary cell.

<https://doi.org/10.7554/eLife.100949.2.sa2>

Reviewer #2 (Public review):

Summary:

This paper presents a very interesting use of a causal graph framework to identify the "root genes" of a disease phenotype. Root genes are the genes that cause a cascade of events that ultimately leads to the disease phenotype, assuming the disease progression is linear.

Strengths:

- The methodology has a solid theoretical background.
- This is a novel use of the causal graph framework to infer root causes in a graph

Comments on revisions:

The authors addressed all of my comments.

<https://doi.org/10.7554/eLife.100949.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Reviewer 1:

(1) The notion of a "root" causal gene - which the authors define based on a graph theoretic notion of topologically sorting graphs - requires a graph that is directed and acyclic. It is the latter that constitutes an important weakness here - it simply is a large

simplification of human biology to draw out a DAG including hundreds of genes and a phenotype Y and to claim that the true graph contains no cycles.

We agree that real causal graphs in biology often contain cycles. We now include additional experimental results with cyclic directed graphs in the Supplementary Materials. RCSP outperformed the other algorithms even in this setting, but we caution the reader that the theoretical interpretation of the RCS score may not coincide with a root causal effect when cycles exist:

“We also evaluated the algorithms on directed graphs with cycles. We generated a linear SEM over $p + 1 = 1000$ variables in $\tilde{X} \cup Y$. We sampled the coefficient matrix β from a

Bernoulli $(1/(p - 1))$ distribution but did not restrict the non-zero coefficients to the upper triangular portion of the matrix. We then proceeded to permute the variable ordering and weight each entry as in the Methods for the DAG. We repeated this procedure 30 times and report the results in Supplementary Figure 3.

RCSP again outperformed all other algorithms even in the cyclic case. The results suggest that conditioning on the surrogate ancestors also estimates the RCS well even in the cyclic case.

However, we caution that an error term E_i can affect the ancestors of $\tilde{X}$ when cycles exist.

As a result, the RCS may not isolate the causal effect of the error term and thus not truly coincide with the notion of a root causal effect in cyclic causal graphs.”

(2) I also encourage the authors to consider more carefully when graph structure learned from Perturb-seq can be ported over to bulk RNA-seq. Presumably this structure is not exactly correct - to what extent is the RCSP algorithm sensitive to false edges in this graph? This leap - from cell line to primary human cells - is also not modeled in the simulation. Although challenging - it would be ideal for the RCSP to model or reflect the challenges in correctly identifying the regulatory structure.

We now include additional experimental results, where we gradually increased the incongruence between the DAG modeling the Perturb-seq and the DAG modeling the bulk RNA-seq using a mixture of graphs. The performance of RCSP degraded gradually, rather than abruptly, with increasing incongruence. We therefore conclude that RCSP is robust to differences between the causal graphs representing Perturb-seq and bulk RNA-seq:

“We next assessed the performance of RCSP when the DAG underlying the Perturb-seq data differs from the DAG underlying the bulk RNA-seq data. We considered a mixture of two random DAGs in bulk RNA-seq, where one of the DAGs coincided with the Perturb-seq DAG and second alternate DAG did not. We instantiated and simulated samples from each DAG as per the previous subsection. We generated 0%, 25%, 50%, 75%, and 100% of the bulk RNA-seq samples from the alternate DAG, and the rest from the Perturb-seq DAG. We ideally would like to see the performance of RCSP degrade gracefully, as opposed to abruptly, as the percent of samples derived from the alternate DAG increases.

We summarize results in Supplementary Figure 4. As expected, RCSP performed the best when we drew all samples from the same underlying DAG for Perturb-seq and bulk RNA-seq. However, the performance of RCSP also degraded slowly as the percent of samples increased from the alternate DAG. We conclude that RCSP can accommodate some differences between the underlying DAGs in Perturb-seq and bulk RNA-seq with only a mild degradation in performance.”

(3) It should also be noted that in most Perturb-seq experiments, the entire genome is not perturbed, and frequently important TFs (that presumably are very far “upstream” and thus candidate “root” causal genes) are not expressed highly enough to be detected with scRNA-seq. In that context - perhaps slightly modifying the language regarding RCSP’s capabilities might be helpful for the manuscript - perhaps it would be better to describe it as an algorithm for causal discovery among a set of genes that were perturbed and measured, rather than a truly complete search for causal factors. Perhaps more broadly it would also benefit the manuscript to devote slightly more text to describing the kinds of scenarios where RCSP (and similar ideas) would be most appropriately applied - perhaps a well-powered, phenotype annotated Perturb-seq dataset performed in a disease relevant primary cell.

We now clarify that Perturb-seq can only identify root causal genes among the perturbed set of genes in the Discussion:

“Modern genome-wide Perturb-seq datasets also adequately perturb and measure only a few thousand, rather than all, gene expression levels. RCSP can only identify root causal genes within this perturbed and measured subset.”

We now also describe the scenario where RCSP can identify root causal genes well in the Introduction:

“Experiments demonstrate marked improvements in performance, when investigators have access to a large bulk RNA-seq dataset and a genome-wide Perturb-seq dataset from a cell line of a disease-relevant tissue.”

Reviewer 2:

(1) The process from health-to-disease is not linear most of the time with many checks along the way that aim to prevent the disease phenotype. This leads to a non-deterministic nature of the path from health-to-disease. In other words, with the same root gene perturbations, and depending on other factors outside of gene expression, someone may develop a phenotype in a year, another in 10 years and someone else never. Claiming that this information is included in the error terms might not be sufficient to address this issue. The authors should discuss this limitation.

The proposed approach accommodates the above non-deterministic nature. The error terms of $\bar{X}$ model factors that are outside of gene expression. We model the relation from gene

expression to Y as probabilistic rather than deterministic because $Y = f_Y(Pa(Y), E_Y)$,

where E_Y introduces stochasticity. Thus, two individuals with the same instantiations of the root causes may develop disease differently. We now clarify this in Methods:

“The error terms model root causes that are outside of gene expression, such as genetic variation or environmental factors. Moreover, the relation from gene expression to Y is

stochastic because $Y = f_Y(Pa(Y), E_Y)$, where E_Y introduces the stochasticity. Two

individuals may therefore have the exact same error term values over $\bar{X}$ but different instantiations of Y .”

(2) The paper assumes that the network connectivity will remain the same after perturbation. This is not always true due to backup mechanisms in the cells. For example, suppose that a cell wants to create product P and it can do it through two alternative paths: Path #1: $A \rightarrow B \rightarrow P$, Path #2: $A \rightarrow C \rightarrow P$. Now suppose that path #1 is more efficient, so when B can be produced, path #2 is inactive. Once the perturbation blocks element B from being produced, the graph connectivity changes by activation of path #2. I did not see the authors taking this into consideration, which seems to be a major limitation in using Perturb-seq results to infer conductivities.

We agree that backup mechanisms can exist and therefore now include additional experimental results, where we gradually increased the incongruence between the DAG modeling the Perturb-seq and the DAG modeling the bulk RNA-seq using a mixture of graphs. The performance of RCSP degraded gradually, rather than abruptly, with increasing incongruence. We therefore conclude that RCSP is robust to differences between the causal graphs representing Perturb-seq and bulk RNA-seq:

“We next assessed the performance of RCSP when the DAG underlying the Perturb-seq data differs from the DAG underlying the bulk RNA-seq data. We considered a mixture of two random DAGs in bulk RNA-seq, where one of the DAGs coincided with the Perturb-seq DAG and second alternate DAG did not. We generated 0%, 25%, 50%, 75%, and 100% of the bulk RNA-seq samples from the alternate DAG, and the rest from the Perturb-seq DAG. We ideally would like to see the performance of RCSP degrade gracefully, as opposed to abruptly, as the percent of samples derived from the alternate DAG increases.

We summarize results in Supplementary Figure 4. As expected, RCSP performed the best when we drew all samples from the same underlying DAG for Perturb-seq and bulk RNA-seq. However, the performance of RCSP also degraded slowly as the percent of samples increased from the alternate DAG. We conclude that RCSP can accommodate some differences between the underlying DAGs in Perturb-seq and bulk RNA-seq with only a mild degradation in performance.”

(3) There is substantial system heterogeneity that may cause the same phenotype. This goes beyond the authors claim that although the initial gene causes of a disease may differ from person to person, at some point they will all converge to changes in the same set of “root genes.” This is not true for many diseases, which are defined based on symptoms and lab tests at the patient level. You may have two completely different molecular pathologies that lead to the development of the same symptoms and test results. Breast cancer with its subtypes is a prime example of that. In theory, this issue could be addressed if there is infinite sample size. However, this assumption is largely violated in all existing biological datasets.

The proposed method accommodates the above heterogeneity. We do not assume that the root causes affect the same set of root causal genes. Instead the root causes *and* root causal genes may vary from person to person. We write in the Introduction:

“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... 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.”

The root causal genes may further affect different downstream genes at the patient-specific level. However root causal genes tend to have many downstream effects so that virtually

every gene expression level becomes correlated with Y . We now clarify this by describing the omnigenic root causal model in the Introduction as follows:

“Finally, application of the algorithm to two complex diseases with disparate pathogeneses recovers an *omnigenic root causal model*, where a small set of root causal genes drive pathogenesis but impact many downstream genes within each patient. As a result, nearly all gene expression levels are correlated with the diagnosis at the population level.”

(4) *Were the values of the synthetic variables Z-scored?*

Yes, all variables were z-scored. We now clarify this in Methods:

“We also standardized all variables before running the regressions to prevent gaming of the marginal variances in causal discovery (Reisach et al., 2021; Ng et al., 2024).”

(5) *The algorithm seems to require both RNA-seq and Perturb-seq data (Algorithm 1, page 14). Can it function with RNA-seq data only? What will be different in this case?*

The algorithm cannot function with observational bulk RNA-seq data only. We included Perturb-seq because causal discovery with observational RNA-seq data alone tends to be inaccurate and unstable, as highlighted by the results of CausalCell. We further emphasize that we do not rely on d-separation faithfulness in Methods, which is typically required for causal discovery from observational data alone:

“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.”

(6) *Synthetic data generation: how many different graphs (SEMs) did they start from? (30?) How many samples per graph? Did they test different sample sizes?*

We now clarify that we generate 30 random SEMs, each associated with a DAG. We used 200 samples for the bulk RNA-seq to mimic a relatively large but common sample size. We also drew 200 samples for each perturbation or control in the Perturb-seq data. We did not consider multiple sample sizes due to the time required to complete each run. Instead, we focused on a typical scenario where investigators would apply RCSP. We now write the following in the Methods:

“We drew 200 samples for the bulk RNA-seq data to mimic a large but common dataset size. We introduced knockdown perturbations in Perturb-seq by subtracting an offset of two in the

softplus function: $\bar{X}_i \leftarrow \text{softplus}(\bar{X}_i - 2)$. We finally drew 200 samples for the control and

each perturbation condition to generate the Perturb-seq data. We repeated the above procedure 30 times.” We also include the following in Results:

“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.”

(7) *The presentation of comparative results (Supplementary Figures 4 and 7) is not clear. No details are given on how these results were generated. (what does it mean “The first column denotes the standard deviation of the outputs for each algorithm?”) Why all other methods have higher SD differences than RCSP? Is it a matter of scaling? Shouldn't they have at least some values near zero since the authors “added the minimum value so that all histograms begin at zero?”*

Each of these supplementary figures contains a 6 by 3 table of figures. By the first column, we mean column one (with rows 1 through 6) of each figure. The D-RCS and D-SD scores represent standard deviations of the RCS and SD scores from zero of each gene, respectively. We can similarly compute the standard deviation of the outputs of the algorithms. We now clarify this in the Supplementary Materials:

“The figure contains 6 rows and 3 columns. Similar to the D-RCS, we can compute the standard deviation of the output of each algorithm from zero for each gene. The first column in Supplementary Figure 7 denotes the histograms of these standard deviations across the genes.”

Many histograms do not appear to start at zero because the bars are too small to be visible. We now clarify this in the Supplementary Materials as well:

“Note that the bars at zero are not visible for many algorithms, since only a few genes attained standard deviations near the minimum.”

(8) Why RCSP results are more like a negative binomial distribution and every other is kind of normal?

All other methods have higher standard deviations than RCSP because they fail to compute an accurate measure of the root causal effect. Recall that, just like a machine has a few root causal problems, only a few root casual genes have large root causal effects under the omnigenic root causal model. The results of RCSP look more like a negative binomial distribution because most RCS scores are concentrated around zero and only a few RCS scores are large – consistent with the omnigenic root causal model. The other algorithms fail to properly control for the upstream genes and thus attain large standard deviations for nearly all genes. We now clarify these points in the Supplementary Materials as follows:

“If an algorithm accurately identifies root causal genes, then it should only identify a few genes with large conditional root causal effects under the omnigenic root causal model. The RCSP algorithm had a histogram with large probability mass centered around zero with a long tail to the right. The standard deviations of the outputs of the other algorithms attained large values for nearly all genes. 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.”

(9) What is the significance of genes changing expression “from left to right” in a UMAP plot? (e.g., Fig. 3h and 3g)

The first UMAP dimension captured the variability of the RCS scores for most root causal genes. As a result, we could focus our analysis on the black cluster in Figure 3 (g) with large RCS scores in the subsequent pathway enrichment analysis summarized in Figure 3 (j). If two dimensions were involved, then we would need to analyze at least two clusters (e.g., black and pink), but this was not the case. We now clarify this in Results:

“The RCS scores of 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 3 (h). We found three exceptions to this rule among the top 30 genes (example in Figure 3 (i) and see Supplementary Materials). RCSP thus detected genes with large RCS scores primarily in the black cluster of Figure 3 (g). Pathway enrichment analysis within this cluster alone yielded supra-significant results on the same pathway detected in the global analysis...”

(10) The authors somewhat overstate the novelty of their algorithm. Representation of GRNs as causal graphs dates back in 2000 with the work of Nir Friedman in yeast. Other

methods were developed more recently that look on regulatory network changes at the single sample level which the authors do not seem to be aware (e.g., Ellington et al, NeurIPS 2023 workshop GenBio and Bushur et al, 2019, Bioinformatics are two such examples). The methods they mention are for single cell data and they are not designed to connect single sample-level changes to a person's phenotype. The RCS method needs to be put in the right background context in order to bring up what is really novel about it.

We agree that many methods already exist for uncovering associational, predictive (Markov, neighborhood) and causal gene regulatory networks. We now cite the above papers. However, the novelty in our manuscript is not causal graph discovery, but rather estimation of root causal effects, detection of root causal genes, and the proposal of the omnigenic root causal model. We now clarify this in the

Introduction:

“Many algorithms focus on discovering associational or predictive relations, sometimes visually represented as gene regulatory networks (Costa et al., 2017; Ellington et al., 2023). Other methods even identify causal relations (Friedman et al., 2000; Wang et al., 2023; Wen et al., 2000; Buschur et al., 2000), but none pinpoint the *first* gene expression levels that ultimately generate the vast majority of pathogenesis. Simply learning a causal graph does not resolve the issue because causal graphs do not summarize the effects of *unobserved* root causes, such as unmeasured environmental changes or variants, that are needed to identify all root causal genes. We therefore define the Root Causal Strength (RCS) score...”

Reviewer 3:

(1) Several assumptions of the method are problematic. The most concerning is that the observational expression changes are all causally upstream of disease. There is work using Mendelian randomization (MR) showing that the opposite is more likely to be true: most differential expression in disease cohorts is a consequence rather than a cause of disease (Porcu et al., 2021). Indeed, the oxidative stress of AMD has known cellular responses including the upregulation of p53. The authors need to think carefully about how this impacts their framework. Can the theory say anything in this light? Simulations could also be designed to address robustness.

Strictly speaking, we believe that differential expression in disease most likely has a cyclic causal structure: gene expression causes a diagnosis or symptom severity, and a diagnosis or symptom severity lead to treatments and other behavioral changes that perturb gene expression. For example, revTMWR in Porcu et al. (2021) uses trans-variants that are less likely to *directly* cause gene expression and instead directly cause a phenotype. However, TWMR as proposed in Porcu et al. (2019) instead uses cis-eQTLs and finds many putative causal relations from gene expression to phenotype. Thus, both causal directions likely hold.

RCSP uses disease-relevant tissue believed to harbor gene expression levels that cause disease. However, RCSP theoretically cannot handle the scenario where Y is a non-sink vertex *and* is a parent of a gene expression level because modern Perturb-seq datasets usually do not perturb or measure Y . We therefore empirically investigated the degree of error by running experiments, where we set Y to a non-sink vertex, so that it can cause gene expression. We find that the performance of RCSP degrades considerably for gene expression levels that contain Y as a parent. Thus RCSP is sensitive to violations of the sink target assumption:

“We finally considered the scenario where Y is a non-sink (or non-terminal) vertex. If Y is a parent of a gene expression level, then we cannot properly condition on the parents because modern Perturbseq datasets usually do not intervene on Y or measure Y . We therefore

empirically investigated the degradation in performance resulting from a non-sink target Y , in particular for gene expression levels where Y is a parent. We again simulated 200 samples from bulk RNA-seq and each condition of Perturbseq with a DAG over 1000 vertices, an expected neighborhood size of 2 and a non-sink target Y . We then removed the outgoing edges from Y and resampled the DAG with a sink target. We compare the results of RCSP for both DAGs in gene expression levels where Y is a parent. We plot the results in Supplementary Figure 5. As expected, we observe a degradation in performance when Y is not terminal, where the mean RMSE increased from 0.045 to 0.342. We conclude that RCSP is sensitive to violations of the sink target assumption.”

(2) A closely related issue is the DAG assumption of no cycles. This assumption is brought to bear because it is required for much classical causal machinery, but is unrealistic in biology where feedback is pervasive. How robust is RCSP to (mild) violations of this assumption? Simulations would be a straightforward way to address this.

We agree that real causal graphs in biology often contain cycles. We now include additional experimental results with cyclic directed graphs in the Supplementary Materials. RCSP outperformed the other algorithms even in this setting, but we caution the reader that the theoretical interpretation of the RCS score may not coincide with a root causal effect when cycles exist:

“We also evaluated the algorithms on directed graphs with cycles. We generated a linear SEM over $p + 1 = 1000$ variables in $\mathbf{X} \rightarrow \mathbf{Y}$. We sampled the coefficient matrix β from a

Bernoulli $(1/(p - 1))$ distribution but did not restrict the non-zero coefficients to the upper triangular portion of the matrix. We then proceeded to permute the variable ordering and weight each entry as in the Methods for the DAG. We repeated this procedure 30 times and report the results in Supplementary Figure 3.

RCSP again outperformed all other algorithms even in the cyclic case. The results suggest that conditioning on the surrogate ancestors also estimates the RCS well even in the cyclic case.

However, we caution that an error term E_i can affect the ancestors of $\mathbf{X}$, when cycles exist.

As a result, the RCS may not isolate the causal effect of the error term and thus not truly coincide with the notion of a root causal effect in cyclic causal graphs.”

(3) The authors spend considerable effort arguing that technical sampling noise in $\mathbf{X}$ can effectively be ignored (at least in bulk). While the mathematical arguments here are reasonable, they miss the bigger picture point that the measured gene expression $\mathbf{X}$ can only ever be a noisy/biased proxy for the expression changes that caused disease: 1) Those events happened before the disease manifested, possibly early in development for some conditions like neurodevelopmental disorders. 2) bulk RNA-seq gives only an average across cell-types, whereas specific cell-types are likely “causal.” 3) only a small sample, at a single time point, is typically available. Expression in other parts of the tissue and at different times will be variable.

We agree that many other sources of error exist. The causal model of RNA-expression in Methods corresponds to a single snapshot in time for each sample. We now clarify this in the Methods as follows:

“We represent a snapshot of a biological causal process using an SEM over $\tilde{X}, Y$ obeying Equation (3).”

We thus only detect the root causal genes in a single snapshot in time for each sample in bulk RNA-seq. If we cannot detect the root causal effect in a gene due to the signal washing out over time as in (1), or if the root causal effect in different cell types cancel each other out to exactly zero in bulk as in (2), then we cannot detect those root causal genes even with an infinite sample size.

(4) While there are connections to the omnigenic model, the latter is somewhat misrepresented. The authors refer to the “core genes” of the omnigenic model as being at the end (longitudinal) of pathogenesis. The omnigenic model makes no statements about temporal ordering: in causal inference terminology the core genes are simply the direct causes of disease.

We now clarify that we use the word *pathogenesis* to mean the causal cascade from root causes to the diagnosis. In this case, the direct causes of the diagnosis correspond to the end of pathogenesis, while the root causes correspond to the beginning. For example, if

$X_1 \rightarrow X_2 \rightarrow Y$, with Y a diagnosis, then X_1 is a root causal gene while X_2 is a core (direct causal) gene. We now clarify this in the Introduction:

“Root causes of disease correspond to the most upstream causes of a diagnosis with strong causal effects on the diagnosis. *Pathogenesis* refers to the causal cascade from root causes to the diagnosis. Genetic and non-genetic factors may act as root causes and affect gene expression as an intermediate step during pathogenesis. We introduce root causal gene expression levels – or *root causal genes* for short – that correspond to the initial changes to *gene expression* induced by genetic and non-genetic root causes that have large causal effects on a downstream diagnosis (Figure 1 (a)). Root causal genes differ from core genes that directly cause the diagnosis and thus lie at the end, rather than at the beginning, of pathogenesis (Boyle et al., 2017).”

(5) A key observation underlying the omnigenic model is that genetic heritability is spread throughout the genome (and somewhat concentrated near genes expressed in disease relevant cell types). This implies that (almost) all expressed genes, or their associated (e)SNPs, are “root causes”.

We now clarify that genetic heritability can be spread throughout the genome in the omnigenic root causal model as well in the Discussion:

“Further, each causal genetic variant tends to have only a small effect on disease risk in complex disease because the variant can directly cause Y or directly cause any causal gene including those with small root causal effects on Y ; thus, all error terms that cause Y can model genetic effects on Y . However, the root causal model further elaborates that genetic and non-genetic factors often combine to produce a few root causal genes with large root causal effects, where non-genetic factors typically account for the majority of the large effects in complex disease. Many variants may therefore cause many genes in diseases with only a few root causal genes.”

We finally add Figure 5 into the Discussion as a concrete example illustrating the omnigenic root causal model:

(6) The claim that root causal genes would be good therapeutic targets feels unfounded. If these are highly variable across individuals then the choice of treatment becomes challenging. By contrast the causal effects may converge on core genes before impacting disease, so that intervening on the core genes might be preferable. The jury is still out on these questions, so the claim should at least be made hypothetical.

We clarify that we do not claim that root causal genes are better treatment targets than core genes in terms of magnitudes of causal effects on the phenotype. For example, in the common cold with a virus as the root cause, giving a patient an antiviral will eliminate fever and congestion, but so will giving a decongestant and an antipyretic. We only claim that treating root causal genes can eliminate disease near its pathogenic onset, just like giving an antiviral can eliminate the viral load and stop pathogenesis. We write the following the Introduction:

“Treating root causal genes can modify disease pathogenesis in its entirety, whereas targeting other causes may only provide symptomatic relief... Identifying root causal genes is therefore critical for developing treatments that eliminate disease near its pathogenic onset.”

We also further clarify in the Discussion that root causal genes account for deleterious causal effects not captured by the diagnosis *Y*:

“We finally emphasize that the root causal model accounts for all deleterious effects of the root causal genes, whereas the core gene model only captures the deleterious effects captured by the diagnosis *Y*. For example, the *disease* of diabetes causes retinopathy, but retinopathy is not a part of the diagnostic criteria of diabetes. As a result, the gene expression levels that cause retinopathy but not the *diagnosis* of diabetes are not core genes, even though they are affected by the root causal genes.”

We do agree that root causal genes may differ substantially between patients, although it is unclear if the heterogeneity is too great to develop treatments.

(7) The closest thing to a gold standard I believe we have for “root causal genes” is integration of molecular QTLs and GWAS, specifically coloc/MR. Here the “E” of RCSP are explicitly represented as SNPs. I don’t know if there is good data for AMD but there certainly is for MS. The authors should assess the overlap with their results. Another orthogonal avenue would be to check whether the root causal genes change early in disease progression.

Colocalization and Mendelian randomization unfortunately cannot identify root causal effects because they all attempt, either heuristically (colocalization) or rigorously (MR), to identify variants that cause each gene expression level rather than variants that *directly* cause each gene expression level and thus make up the error terms. We therefore need new methods that can identify direct causal variants in order to assess overlap.

We checked whether root causal genes change early in disease progression using knowledge of pathogenesis. In particular, oxidative stress induces pathogenesis in AMD, and RCSP identified root causal genes involved in oxidative stress in 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 (Barouch et al., 2007). We therefore expect the root causal genes of AMD to include genes involved in oxidative stress during early pathogenesis. The gene MIPEP with the highest D-RCS score in Figure 3 (d) indeed promotes the maturation of oxidative phosphorylation-related proteins (Shi et al., 2011). The second gene SLC7A5 is a solute carrier that activates mTORC1 whose hyperactivation increases oxidative stress via

lipid peroxidation (Nachev et al., 2021; Go et al., 2020). The gene HEATR1 is involved in ribosome biogenesis that is downregulated by oxidative stress (Turi et al., 2018). The top genes discovered by RCSP thus identify pathways known to be involved in oxidative stress.”

Similarly, T cell infiltration across the blood brain barrier initiates pathogenesis in MS, and RCSP identified root causal genes involved in this infiltration:

“Genes with the highest D-RCS scores included MNT, CERCAM and HERPUD2 (Figure 4 (d)). MNT is a MYC antagonist that modulates the proliferative and pro-survival signals of T cells after engagement of the T cell receptor (Gnanaprakasam et al., 2017). Similarly, CERCAM is an adhesion molecule expressed at high levels in microvessels of the brain that increases leukocyte transmigration across the blood brain barrier (Starzyk et al., 2000). HERPUD2 is involved in the endoplasmic-reticulum associated degradation of unfolded proteins (Kokame et al., 2000). Genes with the highest D-RCS scores thus serve key roles in known pathogenic pathways of MS.”

(8) The available Perturb-seq datasets have limitations beyond on the control of the authors. 1) The set of genes that are perturbed. The authors address this by simply sub-setting their analysis to the intersection of genes represented in the perturbation and observational data. However, this may mean that a true ancestor of X is not modeled/perturbed, limiting the formal claims that can be made. Additionally, some proportion of genes that are nominally perturbed show little to no actual perturbation effect (for example, due to poor guide RNA choice) which will also lead to missing ancestors.

We now clarify that Perturb-seq can only identify root causal genes among the adequately perturbed set of genes in the Discussion:

“Modern genome-wide Perturb-seq datasets also only adequately perturb and measure a few thousand, rather than all, gene expression levels. RCSP can only identify root causal genes within this perturbed and measured subset.”

(9) The authors provide no mechanism for statistical inference/significance for their results at either the individual or aggregated level. While I am a proponent of using effect sizes more than p-values, there is still value in understanding how much signal is present relative to a reasonable null.

We now explain that RCSP does not perform statistical inference in Methods because it is not clear how to define the appropriate cut-off for the RCS score under the null distribution:

“We focus on statistical estimation rather than statistical inference because $\Phi_i > 0$ when E_i causes Y under mild conditions, so we reject the null hypothesis that $\Phi_i = 0$ for many genes if many gene expression levels cause Y . However, just like a machine typically breaks down due to only one or a few root causal problems, we hypothesize that only a few genes have large RCS scores $\Phi_i \gg 0$ even in complex disease.”

(10) I agree with the authors that age coming out of a “root cause” is potentially encouraging. However, it is also quite different in nature to expression, including being “measured” exactly. Will RCSP be biased towards variables that have lower measurement error?

We tested the above hypothesis by plotting sequencing depth against the D-RCS scores of each gene. We observed a small negative correlation between sequencing depth and D-RCS scores, indicating the D-RCS scores are slightly biased upwards with low sequencing depth. However, genes with the largest D-RCS scores exhibited a wide variety of sequencing depths in both MS

and AMD, suggesting that sequencing depth has minimal effect on the largest D-RCS scores. We now explain these results for AMD in the Supplementary Materials:

“Theorem 1 states that RCS scores may exhibit bias with insufficient sequencing depth. The genes with large D-RCS scores may therefore simply have low sequencing depths. To test this hypothesis, we plotted sequencing depth against D-RCS scores. Consistent with Theorem 1, we observed a small negative correlation between D-RCS and sequencing depth ($\rho = -0.16$, $p=2.04E-13$), and D-RCS scores exhibited greater variability at the lowest sequencing depths (Supplementary Figure 8). However, genes with the largest D-RCS scores had mean sequencing depths interspersed between 20 and 3000. We conclude that genes with the largest D-RCS scores had a variety of sequencing depths ranging from low to high.”

We also report the results for MS:

“We plot sequencing depth against the D-RCS scores of each gene similar to the AMD dataset. We again observed a small negative correlation ($\rho = -0.136$, $p < 2.2E-16$), indicating that genes with low sequencing depths had slightly higher D-RCS scores on average (Supplementary Figure 12). However, genes with the largest D-RCS scores again had a variety of sequencing depths. We conclude that sequencing depth has minimal correlation with the largest D-RCS scores.”

| (11) Finally, it's a stretch to call K562 cells "lymphoblasts." They are more myeloid than lymphoid.

We now clarify that K562 cells are undifferentiated blast cells that can be induced to differentiate into lymphoblasts in Results:

“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 undifferentiated blast cells that can be induced to differentiate into lymphoblasts, or the precursors of T cells and other lymphocytes.”

<https://doi.org/10.7554/eLife.100949.2.sa0>