Title: The Polygenic Score - Rare Variant Causal Pivot: A Conditional Approach to Discovery in Complex Disease Genetics

Authors and Affiliations:

Chad Shaw[1,2,4]*, Christopher Williams[3], Tao Tao Tan[1], Nicholas Di[2], Joshua Shulman[1,4] and John W. Belmont[1,3]

1. Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas
2. Department of Statistics, Rice University, Houston, Texas
3. Genetics & Genomics Services, Inc., Houston, Texas
4. Department of Neurology, Baylor College of Medicine, Houston, Texas

*Correspondence: cashaw@bcm.edu

Abstract

We develop a structural causal methodology to analyze genetic heterogeneity in complex diseases. Our method, the Causal Pivot (CP), exploits one known cause to detect the contribution of a second suspected cause. We employ polygenic scores (PRS) as a known cause; we consider selected rare variation (RV) and RV ensembles as candidate causes. We condition on disease to determine a method that incorporates outcome induced association. We derive a conditional maximum likelihood procedure for both binary and quantitative traits, and we develop a likelihood ratio test (CP-LRT). We demonstrate the CP-LRT's strong power and robustness to misspecification of the RV rate compared to alternatives.

We then use data from the UK Biobank (UKBB) to examine performance. As demonstrations, we consider hypercholesterolemia (HC, LDL-c direct ≥ 4.9 mmol/L; n_c=24656), breast cancer (BC, female, ICD10 C50; n_c=12479) and Parkinson's Disease (PD, ICD10 G20, n_c=2940). For PRS we use UKBB data; for RV we select ClinVar pathogenic and likely pathogenic variants in known disease genes. For HC we focus on *LDLR,* for BC we focus on *BRCA1*, and in PD we consider *GBA*. Using CP-LRT we detect RV signals for all three demonstration diseases. Cross-disease analyses serve as comparators. We extend our CP-LRT approach using matching and inverse probability weighting to address potential confounding driven by ancestry. Finally, we present a CP oligogenic load analysis in PD using lysosomal storage. Taken together our work demonstrates the versatility of the PRS-RV CP as a method of inference and discovery in complex disease genetics.

Introduction

Causal heterogeneity, where root causes of disease vary among affected individuals, is a signature characteristic of complex disease. Affected individuals may reach the disease phenotype through distinct etiologies which may be influenced by alternative genetic factors. Although the causes of disease vary, they ultimately lead to a shared pathophysiology or disease state. Moreover, complex disease genetic architectures are known to vary from monogenic to highly polygenic.[1,2] Oligogenic inheritance, environmental phenocopies, and gene X environment interactions are also thought to contribute to complex disease etiologies. This causal heterogeneity impacts both prevention and management because distinct causes may require different approaches for effective interventions or risk mitigation.[3]

Genome wide statistical association studies (GWAS) in complex disease are well developed for both rare and common DNA variants[4,5], but integration across the allele frequency spectrum remains challenging. Moreover, non-familial association approaches are typically case-control designs and may not be applicable to all settings.[6,7] Although GWAS has tools for both rare and common variants, almost all studies are marginal analyses that treat each genetic locus as a separate testing unit; results are summarized with Manhattan plots. The marginal approach avoids confronting heterogeneity with respect to alternative genetic factors differentially present or absent in individuals. Consequently, these methods do not classify individuals into causal groups or subtypes. Resolution of heterogeneity is desirable to avoid overgeneralizing results across different subpopulations and to identify which specific causal mechanism or pathway is potentially responsible in individual cases.[8]

Genetic variants -- while exhibiting allelic heterogeneity, pleiotropy and potential confounding with ancestry and environment -- are explicitly considered to be causal factors that drive downstream disease relevant processes. Consequently, methods developed for causal inference and structural causal modeling (SCM) in fields such as epidemiology and econometrics [9,10] have been adapted to problems in genetics. The most successful example is

Mendelian randomization (MR)[11-14], which uses genetic variants as instrumental variables to evaluate potential causal biomarkers and endophenotypes for common diseases. We reasoned causal analysis methods could have broader applications in genetics[15-19] and could offer a potential approach to reframe inquiry into heterogeneity in complex disease genetics.

We hypothesized that we could adopt SCM as a framework to consider multiple genetic factors differentially present or absent in individuals to examine the underlying mechanistic heterogeneity of complex disease. This approach has the goal of deconvolving patient populations into causal groups or subtypes; it has the additional goal of providing individual level insights following the diagnostic process where evaluating and eliminating alternative potential causes is fundamental to personalized medicine.

We develop a statistical approach which we call the Causal Pivot (CP) which can be applied in cases only or a case-control context. Our CP is based on a structural model of disease as well as the well-known collider bias concept[20-22] in which conditioning on an outcome induces a synthetic correlation between otherwise independent causes. Under these conditions, if we observe the disease status of individuals and condition on their outcome state – either case or control —then, within those outcome groups, independent causal factors will have induced correlation in mathematically well defined patterns. Importantly, model factors will show no induced correlation if the factors are unconditionally independent and not mutually causal. We view outcome induced association between distinct causal factors as a feature to resolve heterogeneity and to drive discovery instead of as a source of bias in the research process.[23-25]

To develop CP we reasoned that in most research and diagnostic scenarios investigators have some prior knowledge -- although incomplete -- about causal factors and the strength of their relationship to outcomes. By exploiting this known information together with the induced correlation between a known factor and a candidate cause conditional on outcomes we reasoned it might be possible to pivot from known causal factors to discover novel causes.

Moreover, this approach could also enable discovery in cases-only analysis because the induced correlation would be present without the need for comparative control individuals.

We proceed from the premise that GWAS has uncovered strong evidence for genetic loci driving disease and that in aggregate the observed genetic associations represent causal factors. We hypothesized that we could recruit polygenic risk scores (PRS) as a genome wide representation of genetic causal contribution to complex disease outcomes. Moreover, because variants identified in GWAS are frequently common and because PRS are frequently restricted to common variants above a given minor allele frequency cutoff (i.e. MAF >0.01), we can take MAF restricted PRS as representative of common genetic variation's contribution to disease which excludes rare variation (RV, MAF < 0.01). Therefore, we chose to treat the PRS as a CP to search for new causes by incorporating induced correlation between alternative disease drivers and PRS conditional on outcomes. We focus attention on RV or RV ensembles as candidate causes (Figure 1). The general concept that low PRS can be used to prioritize individuals for RV sequencing or to classify them as having monogenic disorders has been noted[26,27], and more generally, the use of colliders or v-structures to infer causal graphs is well known.[28-30] However, more detailed statistical analysis is still needed to elevate and generalize the procedure, and recognition of the overarching paradigm for PRS - RV causal opposition with respect to complex traits has been lacking. Importantly, prior methods have been non-parametric, lacking explicit parameter estimation and lacking distinction between main effects and interaction effects between PRS and RV. Moreover, non-parametric methods lack interval estimation to quantify uncertainty in results. Moreover, the lack of statistical formalism has hampered clarity on performance of the method and its extension to address confounding by ancestry as well as alternative models such as pathway load.

We developed a highly general statistical framework for the CP using SCM and Bayes rule. This approach leads naturally to a likelihood framework, maximum likelihood estimation, confidence intervals, and to a CP likelihood ratio test (CP-LRT). The same mathematical

structure applies in the context of both binary and quantitative traits. We performed power analyses and compared CP-LRT approach to alternatives [26,27] . Importantly, the CP-LRT derives power from both the disease rate change in RV as well as the conditionally induced dependency between RV and PRS. This insight leads to demonstration of the robust properties of the CP-LRT in terms of sensitivity and specificity. We then demonstrate the performance of CP-LRT in complex disease examples of breast cancer (BC), hypercholesterolemia (HC) and Parkinson's disease (PD using the UK Biobank (UKBB). We then extend CP to address confounded structures for PRS-RV association potentially driven by ancestry. We demonstrate two approaches to address ancestry confounding using both matching and inverse probability weighting. Finally, we develop an RV pathway burden test to demonstrate the possible utility of cases-only CP to investigate oligogenic models for pathway contribution to complex disease using lysosomal storage in PD as an example. Our results demonstrate the generality and potential of CP as an approach to causal discovery and to address heterogeneity in complex disease genetics.

Methods

Causal Models and Graphs

When structural modeling is equipped with a counterfactual interpretation it is also called causal analysis or structural causal modeling. Our methods are developed through structural causal modeling, and we adopt this counterfactual interpretation. In our context the counterfactual assertion is that **but for** the presence or absence of genetic exposures the potential health outcome of an individual could have been different. In the structural framework directed acyclic graphs are used to express model assumptions where directed arrows depict stochastic dependencies and the lack of a directed path between nodes indicates unconditional independence.

We consider simple models with three or four nodes and two or four edges, respectively (**Figure 1 a, d** and **Figure 4 a,b,c**). The disease outcome variable is denoted **Y**. This outcome represents the phenotypic status, for example as recorded in the UKBB metadata by ICD-10 codes or biomarker measurement information. This outcome variable can be binary or continuous. In our work we focs on two genetic exposures: a PRS -- a continuous variable -- which we will denote as **X** representing the combined causal contribution of common variation (MAF >0.01), and a binary RV variable denoted as **G** that represents the presence of a strongly acting RV (**Figure 1**). This binary RV represents the state that an individual is positive for any of an ensemble of strongly acting RV that in themselves may be vanishingly rare or singleton but in aggregate can have frequency approaching 0.001 or larger in a population or study sample. These **X** and **G** are depicted as nodes with edges incident to the **Y** outcome (**Figure 1 a**). Later we extend this three-node model to a four-node graph (**Figure 4 a**) and we let the PRS and RV variables be coordinately influenced by ancestry denoted **A** with arrows incident to **X** and **G;** thereby **A** acts as a common cause in the four node, four edge model.

Causal Pivot

When two causes are incident to the same effect that effect is called a collider for the two causes. Conditioning on the collider outcome induces an observational – but not causal – correlation between the distinct causes. This situation is a well-known source of bias in epidemiologic studies because researchers may unknowingly condition on a collider structure and report the induced association between variables **X** and **G** as causal.

The CP exploits collider induced correlation as a source of signal rather than noise. When one causal variable incident to a collider has a **known effect on the outcome** – such as the effect of the **PRS** on **Y**, then the outcome-conditionally-induced **RV-PRS** correlation can be used to estimate and to test the causal **Y~RV** relationship. The Null Hypothesis is an absence of the **Y~RV** relationship, and in this situation no induced correlation will appear between **X** and **G**

conditional on **Y**. As we show in the Mathematical Supplement, the CP is generally applicable and not limited to genetics.

Conditional Analysis

Here we present an overview of the mathematical derivation of the CP approach; more details are in the Mathematical Supplement. In what follows we refer to the variables PRS as **X**, RV as **G** and disease outcome **Y.** The CP derives from manipulation of the probabilistic factorization determined by the graphical model (**Figure 1A**). The joint density under the model is:

$$f_\theta(x, g, y) = f_{\theta_R}(y \mid x, g) f_{\theta_X}(x) f_{\theta_G}(g) \tag{1}$$

In the expression $f_\theta(\cdot)$ represents the respective probability density functions for their arguments, and the bar symbol expresses conditioning on variables to the right-hand side of the bar. The parameter vector θ represents the unknown parameters governing the probabilistic structure of the system; for clarity we emphasize that the parameter θ is comprised of the parameters θ_R that determine the forward conditional relationship between **Y** and **X**,**G** and the separate parameters θ_{XG} determine the distributions of **X** and **G** – which in the simple model are exogenous and independent.

The Causal Pivot procedure follows from application of Bayes' rule to the probabilistic factorization implied by the graphical model. In notation applying Bayes rule to (1):

$$f_\theta(g \mid x, y) = \frac{f_\theta(y \mid x, g)}{E_G(f_\theta(y \mid x, G))} f_\theta(g) \tag{2}$$

For **G** representing RV status, the expression (2) states that the conditional probability for a sampled individual to bear an RV – denoted by **G=1** – or to not have an RV -- denoted **G**=0 -- is given by the ratio of the forward conditional probability of their phenotype **Y** given PRS **X**

and **G=b** divided by the average probability of **Y** integrating out **G** multiplied against the prior probability of **G=b**. This factorization is general and does not depend on the model specification of **f(y | x,g)** or **f(x)** or **f(g)**; moreover, the factorizations in (1) and (2) do not depend on a univariate restriction on **X**, **G** or **Y**, nor does it depend on the binary character of **G** or the distributional character of **X**.

Model Specification

It is useful to provide a parametric model in order to perform analyses, make calculations, enable parametric testing and to perform power comparisons. Non-parametric approaches also apply, but parametric modeling provides both mathematical insight and yields practical results such as confidence intervals. We adopt a generalized linear model framework for the forward distribution of **Y** given **X** and **G.** The general linear model framework is encompassing of both logistic regression in the case of binary outcomes as well as linear regression for quantitative traits. The general expression for the forward conditional expectation of **Y** given **X** and **G** is:

$$h\left(E\left(Y \mid X, G\right)\right) = \alpha + \beta \cdot X + \gamma \cdot G + \eta \cdot X \cdot G \tag{3}$$

Examples for the function h() are the logit or log-odds for binary outcomes, such as for a binary trait like BC; the identity function can be used for continuous **Y**, such as in the case of LDL measurements. We assume the parameters α and β are known or estimable from population data; we also assume knowledge of a baseline frequency of RV, set at ω. In many cases the estimate of the **X**, **Y** relationship is reported. If not, the easiest way to estimate it is to apply a forward model to unconditional **Y** including **X** and exclude **G.** Although the α parameter depends on **G** we argue that it is negligible because **G** is rare. We adopt this approach to determine α and β in our demonstration work. The frequency of rare variation is estimable from population tabulation such as GNOMAD or from UKBB summary information.

Likelihood Analysis Conditional on Y

To examine the properties of the CP we constructed a test procedure for the possible effect of RV using the factorization in (2) to determine a likelihood function for observed data conditional on **Y=a.** We executed this for both binary **Y** and for Y>δ for continuous **Y**. In the case of discrete Y, we have:

$$\begin{aligned} L(\gamma, \eta \mid X, G, Y = a) &= \prod_{i=1}^{n} f_\theta(g_i \mid x_i, y_i = a) \\ &= \prod_{i=1}^{n} \frac{f_\theta(y_i = a \mid x_i, g_i)}{E_G(f_\theta(y_i \mid x_i, G))} f(g_i) \end{aligned}$$

(4)

Taking logarithms and emphasizing that the unknown elements of θ $\tau\eta\alpha\tau$ concern the impact of the RV on outcome are γ,η:

$$\begin{aligned} l(\gamma, \eta \mid X, G, Y = a) &= \sum_{i=1}^{n} l(\gamma, \eta \mid g_i, x_i, Y_i = a) \\ &= \sum_{i=1}^{n} l(Y_i = a, \gamma, \eta \mid x_i, g_i) - log\left(E_G(f(Y_i = a, \gamma, \eta \mid x_i, G))\right) + l(g_i) \end{aligned}$$

(5)

As shown in the Mathematical Supplement this likelihood function can be maximized in the variables γ and η to produce maximum likelihood estimates; this analysis is performed by differentiating the log-likelihood with respect to γ and η and solving for the root where the derivative is zero. Confidence intervals are obtained using Fisher Information as using the observed means of the squared derivatives of the log-likelihood. We performed this analysis for both logistic and a liability model for a continuous trait (see Supplemental Materials). For the continuous trait we consider the circumstance where the observable sample comprises only individuals with a sufficiently extreme quantitative trait Y>δ where individuals in this class are deemed cases and individuals not in the class could be deemed controls; we suppose we have

access to the trait values for the cases. Importantly, we note that the CP conditional analysis can be performed on cases only, controls only, or on both cases and controls taken together but where the analysis is performed conditionally in both groups. The availability of appropriate research subjects determines which is pursued. In our investigation we focus on cases only analyses.

Likelihood Ratio Test

As mentioned, the likelihood approach determines maximum likelihood estimates (MLE) by solving for the roots of the derivatives of log-likelihood equations treating these as functions of the unknown parameters. A likelihood ratio test can be constructed by plugging in the maximum likelihood estimates into the log-likelihood function to compute a test statistic. The Null Hypothesis is that the parameters γ and η are 0. Under the Null Hypothesis the asymptotic distribution of : $-2 * l(0,0 \mid X,Y,G) - l(\gamma, \eta \mid X,Y,G)$ is a chi-square on two degrees of freedom. The procedure considers the influence of the RV – both individually and in its potential interaction with the PRS. Additional details are provided in the Mathematical Supplement.

Logistic Model

The logistic model considers the disease outcome to be binary where the log-odds of disease is linear in the explanatory variable. We additionally treat RV as a binary variable. In the initial treatment there is no additional stochasticity in the conditional distribution of the outcome **Y** given **X** and **G** except for the coin-flipping variability of a Bernoulli trial; such extensions are possible but out of scope in this paper. The log-likelihood analysis conditional on **Y=a** produces four different sums: the sum of contributions for **Y=0** where **RV=1**, **Y=0** where **RV=0**, **Y=1** where **RV=1**, and **Y=1** where **RV=0**. Breaking the likelihood calculation into these four parts is both computationally practical and useful for analysis. In cases only analysis there is no contribuition from groups where **Y=0.** The details of the log-likelihood analysis and derivative derivation appear in the R markdown code and in the Supplement.

Liability Model

The Causal Pivot liability model considers conditioning on the outcome Y>δ for a quantitative trait that is considered to be conditionally distributed with a mean as in (3) and where additional exogenous noise is Gaussian. The approach generalizes to non-Guassian noise, but the Gaussian situation is both canonical and illustrative. Here the model considers the outcome variable Y | I(Y>δ), x, g. The forward model has a distribution:

$$f(y \mid I_{Y>\delta} = 1, x, g) = \begin{cases} \frac{f(y|x,g)}{1-F(\delta|x,g)} & Y > \delta \\ 0 & Y \leq \delta \end{cases}$$

(6)

For situations where the outcome value is less than δ we have:

$$f(y \mid I_{Y>\delta} = 0, x, g) = \begin{cases} \frac{f(y|x,g)}{F(\delta|x,g)} & y < \delta \\ 0 & y \geq \delta \end{cases}$$

(7)

To perform analysis with this liability model we use the general expression in (2) that becomes expression (5) after applying the log-likelihood transformation. As with the logistic model there are four configurations: where the outcome Y< δ and RV=1, where Y<δ and RV=0, where the Y>δ and RV=1, and where Y>δ and RV=0.

Power Analysis

To perform power analysis, we built a forward simulation procedure to generate outcome data under both the logistic and liability framework. For each scenario, we simulated RV as a binary variate with frequency 0.001. For PRS we generated unit normal outcomes. We fixed β according to values consistent with real data. We let γ and η range across many possible

values. For logistic regression we set α=-2.2 and β =0.5, we set γ and η to values collection of alternatives. We generated **Y** outcomes according to the inverse transformed log-odds. We then applied the likelihood expression in (5) to obtain the MLE, and we compute the likelihood ratio test statistic. We performed 1000 simulations at each model parameterization, and we recorded outcomes. To determine power, we calculated the proportion of simulation runs where the observed p-value was less than 0.05. For the liability model we generated outcome variables in a parallel fashion, first generating X ~N(0,1) and G~ber(0.001) and fixing the parameters α, β and a grid of alternative **RV** effect sizes. We used the traditional unconditional case-control forward regression as a comparator for CP power. This analysis considers a logistic regression for **Y** given **X** and **G** and a linear model for **Y** given **X** and **G** in the quantitative outcome situation.

Cases Only Analysis and Case Control Conditional Analysis

Our CP procedure can be applied to cases-only data, to controls only data or to case-control data taken together. In all of these situations, data analysis proceeds according to a conditional analysis using the fundamental factorization in (2) and to application of the maximum likelihood procedure in (5) followed by a likelihood ratio test. We focused on comparison of cases only CP analysis against the traditional forward regression approach and to alternative cases-only analyses.

Comparator Models and Robustness Analyses

The CP-LRT benefits from the full information in the stochastic system as described by the structural causal graph and as encapsulated in the conditional likelihood function. This conditional likelihood gives the expected rate change in G given both X and Y. We considered two comparator non-parametric models to our CP-LRT . First, we consider a Z-test procedure where we assume the frequency of RV status in the source population –deemed ω -- is exactly known. Under this scenario if the RV status has no effect on outcome, the expected number of

RV in the outcome group **Y=a** would be binomial with expectation $N_{Y=a}\omega$ and variance $N_{Y=a}\omega(1-\omega)$. We take **Z** to be the difference between the observed RV+ disease individuals ($N_{Y=a,RV+}$) and the expected under H_0 standardizing by square root of variance. This comparator focuses on the rate change of a functional variant state in the disease population. Alternatively, prior work proposed a Wilcoxon test on the PRS values among RV+ and RV- individuals in Y=a. The Wilcoxon is a non-parametric test for a shift in means between two populations, in this case RV+ and RV- conditional on Y=a.

Robustness Analyses

We consider the impact of the precision with which ω is known. We examine the type 1 error when **G** has no effect and scrutinize the impact of underspecification of the rate ω in the general population. The type 1 error is the rate at which the CP-LRT and Z-test reject H_0 and falsely identify an effect of **G** when there is none. These analyses are performed across a range of underspecification of the RV rate which we deem ω^* ; the true simulation parameter values in these analyses are α=-2.2, β=0.6,ω=0.001, γ=0,η=0.

We then consider the power impact of overspecifying the RV rate. We examine this situation in the context where α=-2.2, β=0.6, ω=0.001,γ=2.2, η=0. In this case there is an RV effect, but oversetting the value of ω^* makes it more difficult to detect. We examine a range of values ω^* up to and beyond the conditional frequency of RV of 0.0045.

UKBB Data

We analyzed human data from cohorts and controls drawn from the UKB under an approved application (Project 98786) titled “Genetic Heterogeneity in Diseases with Complex Inheritance”. We used hypercholesterolemia (LDL direct | Instance 0 ≥ 4.9 mmol/L; Data-Field 30780), breast cancer (ICD-10 C50; Data-Field 40006), and Parkinson’s Disease (ICD-10 G20; Data-Field 131022) as disease models with expected positive findings. Cases were filtered

based on inferred European ancestry, availability of exome and genotyping data, female-only for breast cancer, and availability of a phenotype-specific polygenic risk score. The final samples included: hypercholesterolemia (HC cases 24656; controls 347889); breast cancer (BC, cases 12479; controls 198311); and Parkinson's (PD, cases 2949; controls 388198). As this study exclusively utilized de-identified data from the UKBB, it qualifies for exemption from ethical review under the guidelines for research involving non-identifiable human data.

PRS

We used PRS made available through the UKBB (Field ID 26220 Standard PRS for BC; 26250 Standard PRS for low density lipoprotein cholesterol; and 26260 Standard PRS for PD).[31]

Exome Variant Extraction and Classification

We used PLINK2 to read genotypes from the UKBB exome BGEN files (Data Field 23159) and extract variants within our target gene lists (see bedfile https://github.com/chadashaw/causal-pivot/) for the disease-defined cohort samples. The extracted genotype files were converted to Matrix Market format using a custom python script. Variants were, then, annotated with OpenCRAVAT; annotators included those pulled from the OpenCRAVAT respository (gnomad3, revel, cadd_exome) as well as custom annotators (tcg_clinvar, AlphaMissense, ESM1b) produced from publicly available data. We exclude variants that were absent (homozygous-ref) across all samples.

We focused on three example genes for the demonstration diseases: *BRCA1* for BC, *LDLR* for HC and *GBA* for PD. We used the OpenCRAVAT annotations to classify rare deleterious and/or pathogenic variants within these 3 genes. Rare deleterious variants were those with a gnomAD v3 Non-Finnish European allele frequency of less than 0.1% and were, functionally, either stop-gain or frameshift. Pathogenic variants were ClinVar "Pathogenic" or "Likely Pathogenic" without conflicting "Benign" annotations. Variants not meeting either of these criteria were considered negatives.

We evaluated as RV positive those individuals with at least one alternate allele for any rare deleterious or pathogenic variant within each gene. To ensure independence between our two causal factors (RV status and PRS) we further excluded all variants from each gene strongly associated with the PRS for that gene's demonstration disease. We used logistic regression to estimate the statistical significance of any correlation between individual variants and their associated PRS and removed those with an effect p-value of less than 0.05.

We extracted relevant cohort metadata (age, sex, disease indicators, LDL, and ancestry components) from the UKB cohort browser.

Controlling for Ancestry

We consider potential confounding between PRS and RV arising from shared Ancestry, denoted A (**Figure 1B**). We adopt two approaches to address this situation: matching and inverse probability weighting. With both methods we focus on the conditional outcome where Y=1, cases only analysis. We expect that if RV influences Y then a negative association will appear between RV and PRS such that affected individuals with RV=1 will have an average lower PRS than affected individuals with RV=0.

To perform the matching, samples in each of the three cohort groups were differentiated based on positive disease (case) status. A distance matrix between all samples was calculated using the `distances` package in R. In the absence of eigenvalues for the available PCs in UKBB an exponentially decreasing weighting function (scaled to unity) [decay = e^{r*5}, weights = decay / sum(decay) **r = 0.637**] was used to weight distances for each PC in descending order. RV+ cases were matched to the closest (k=2) RV- cases using the calculated distances. For each RV+ case, the difference between the observed PRS value and the mean PRS value for the matched RV- cases was calculated. An overall test statistic was calculated by taking the average of these values over all RV+ cases. Permutation testing was performed using the RV+

status as the treatment variable. A 1-tailed permutation test of means was then performed using the observed test-statistic against the permutation distribution.

We also performed an analysis using inverse probability weighting. Briefly, inverse probability weighting is a general method for normalizing outcome data in causal analysis to account for unequal representation of the causal exposure that correlates with an observable common cause. In this situation the outcome variable is taken to be the PRS among the affected individuals and the exposure is the binary RV; the common cause is ancestry. We use the random forest method to perform a machine learning predictive analysis for RV positivity using the ancestry principal components as predictors in each disease cohort. The outcome of random forest is a probability of RV=1 for each observation. We then computed the inverse probability weighted sum of PRS among RV+ and subtracted the sum of one minus the probability of RV of the respective in the RV- cases. We again use permutation assignment of the RV status to determine a null distribution and a one sided p-value.

Test of RV Load in Individuals

We tabulated a count of ClinVar pathogenic variants in a pathway and considered a count variable as the candidate cause for the CP against PRS; in this approach the count is total RV variants in a pathway using the same rule as for our binary RV analyses. For this oligogenic analysis we focus on the lysosomal storage pathway in PD. We enumerated variants in 54 lysosomal storage genes that were also present in UKBB. We then counted the number of variants in each individual. To examine association of this count variable with PRS conditional on disease we use Poisson regression to consider the average RV count as a function of PRS. Briefly, in Poisson regression the logarithm of the rate of RV is modeled as a linear function of the PRS. We used the LRT for Poisson regression as implemented in R to determine p-values.

Results

Comparative power analyses

We hypothesized our Causal Pivot LRT (CP-LRT) would have promising statistical properties to discover RV causal candidates by exploiting a PRS as a known cause. The model assumes the known factor is previously established and has a known effect size. The candidate cause is a binary variable representing the presence of at least one functional RV. This variable is assumed to have a known sparse rate of occurrence in the general population, and we use the value 0.001. Simulation-based power analyses presented in **Figure 2a** show the performance of the CP-LRT in the context of a binary outcome trait considering the possibility of both a main effect of RV as well as a PRS-RV interaction. In all simulations the forward logistic regression using the full data of both cases and controls has the highest power; this test is using approximately an order of magnitude more subjects because n=500,000 for case-control, vs n_c of approximately 50,000. Importantly, using only cases – approximately 10% of the population – achieves 90% of the power of the case-control analysis for the CP-LRT when the effect of the RV is at least as large as the effect of the PRS. In the situation of cases-only analyses with no interaction effect, the Z-test looking for overrepresentation of RV+ individuals has the highest power. When the RV main effect is sufficient and there is an interaction between RV and PRS the CP-LRT has superior power over the Z-test. The CP-LRT has superior power over the previously proposed Wilcox test (purple curve) when there is no interaction effect.

For the liability model (**Figure 2b**) we simulated according to a linear model for continuous quantitative outcomes. We modeled our analysis on the example with parameters α=0, β=0.3, ω=0.001, and we used a cutoff value of δ=1.13,corresponding to the 92%-ile of the trait distrubtion, consistent with the LDL cholesterol cutoffs. Conditional cases-only analysis corresponds to the situation where observed samples are restricted to those with trait values over a cutoff. In this situation the CP-LRT has power that is stronger than the Wilcox test when

there is no interaction effect. When the interaction effect is strong the Wilcox test has power of unity. The forward linear regression model using both case and control data has the strongest power, and the Z-test also has high power.

We also perform robustness analysis focusing on the binary trait situation in the cases-only circumstance. In this situation analyses for the Z-test and the CP-LRT rely on prior information concerning RV variant frequency. Therefore, we explore the impact of errors in the specification of this frequency. **Figure 2c** considers the situation where there is no RV effect, with both γ=0 and η=0. Results show that the CP-LRT hast lower Type 1 error rate compared to the Z-test, as indicated by the cyan Z-test curve dominating the red CP-LRT curve. The CP-LRT retains a type 1 error of 0.05 when the RV rate is misspecified by 10%, from 0.001 down to 0.0009.

Conversely, the rate parameter for the RV might be falsely specified above its true value. Figure **2d** shows this situation when the RV effect is set to $\alpha=-2.2$, with γ=2.2 and the PRS parameter is β=0.6 – values driven from the UKBB data for BC. In these cases, the CP-LRT has higher power to detect an effect while the Z-test power is sharply reduced, as indicated by the red CP-LRT curve dominating the Z-test cyan curve. Interestingly, the CP-LRT retains a power of 75% when the RV frequency is set to the expected expected RV frequency among cases of 0.0045; in this circumstance the Z-test power is only 0.05.

Application to UKB data

To assess relevance of CP in real data we used UKB exomes and PRS. We used well studied complex traits with both valid standard PRS as well as established single gene contributors. The latter allowed us to recruit ClinVar annotations to select disease causing rare variants as positive controls. Disease cohorts and variant selection are as described in the Methods.

Cross Cohort Comparisons

We employ a cross-cohort comparative approach. We extract affected individuals in each disease population and their respective PRS value as well as an RV determination mining exome data for qualifying RV (see Methods). For visualization of the relative rate of RV against PRS we split the respective cohorts into tertiles by their PRS. For each PRS tertile in each cohort we counted the rate of RV in each PRS tertile bin; this analysis is repeated for each of the three example genes: *BRCA1*, *LDLR* and *GBA*, respectively. We perform statistical testing by application of our CP-LRT treating PRS quantitatively (**Figure 3a**). For each gene we perform a CP-LRT for the rate of RV vs PRS by cohort. The diagonal of the plot are highly significant ($p < 1x10^{-6}$) indicating that the CP-LRT found a significant negative association where the matching gene RV state is tabulated against its disease specific PRS. Cross-cohort analyses are mostly not significant, indicating that RV status for *BRCA1* and *GBA* are not associated with either PD state or BC state, respectively. The gene *LDLR* shows weak but significant CP-LRT association with both BC and PD. Supplemental Figure 2 shows the MLE estimates of the RV and interaction parameters; the point estimates and confidence intervals are provided. The point estimates using the CP MLE strongly agree with the results from case-control analyses.

Controlling for the effect of Ancestry

Ancestry may confound the PRS-RV relationship by acting as a common cause for an individual to have a functional RV as well as a high or low PRS. We call this scenario the diamond graph (**Figure 4 a-c**). We use both matching and inverse probability weighting to examine for signal while controlling for this possible confounding effect.

For matching we use the ancestry principal components to identify neighborhoods of affected individuals. In each neighborhood we take the difference between the PRS of the RV+ and RV- affected individuals, and then average the PRS differences. The results show that cases the RV+ individuals have on average significantly lower PRS scores than RV- individuals after controlling for ancestry (**Figure 4d**).

Propensity scoring and inverse probability weighting is an alternative approach to control for confounding. We treat RV status as a binary exposure, and we construct a model to predict the chance of this exposure using Ancestry information from UKB population PC values. We use random forests to predict the probability of RV+ status using the first 10 principal components. We then compute the IP-weighted differences in PRS between RV+ and RV- individuals. The results again show that PRS values are significantly lower in RV+ compared to RV- individuals (**Supplemental Figure 1**).

Generalization to pathway burden scores

We considered the lysosomal storage pathway in the context of PD as an example of pathway load analysis in the context of the CP. The lysosomal storage pathway is the system that supports recycling of macromolecules in the brain. There are many lines of evidence that point to lysosomal storage as a central pathway in neurodegeneration, and it plays a key role in PD. We established a set of 54 human lysosomal storage pathway genes (**Supplemental table**). We tabulate for each individual the number of qualifying RV in the lysosomal storage pathway for each individual. We then compute the mean PRS of individuals with disease stratifying on the number of pathway RV. We then use a Poisson Regression to examine the association of the PRS with pathway burden of RV. The results are presented in **Figure 5a**. Among the affected individuals there is a downward trend in PRS with increasing pathway burden. To emphasize the combinatorial sparsity of the data we present the count of two-variant combinations among cases **Figure 5b**.

Discussion

To date, the general principles of causal inference and structural causal models (SCM) have been applied in complex trait genetics using Mendelian Randomization (MR)[14,16] which exploits DNA variants as instrumental variables. Importantly, MR methods are typically used to investigate single mechanisms that flow through a candidate biomarker and are not used to

differentiate individuals based on multiple independent causes or causal classes. Given the success of MR, it is likely that SCM can be used even more broadly to model genetic effects, suggesting several new avenues for addressing common diseases and other complex traits.[32]

The study of independent causes, in classical genetics called locus and allelic heterogeneity, is central to understanding the genetic architecture of monogenic disorders.[33] Parallel concepts of causal heterogeneity are needed for complex traits where current methods focus on the average causal contribution of variants rather than their operation in potentially heterogeneous subgroups.[34] There are already well-known examples in certain complex traits in which the total population of affected individuals includes both those with high polygenic risk along with individuals affected by monogenic forms. In this work, we have developed and evaluated an approach to address this form of causal heterogeneity using a general SCM framework. By examining the possible conditional relationship between a known cause and a distinct candidate cause conditional on the phenotypic outcome we demonstrate that one can establish their independent contributions. Moreover, we show that this approach applies in cases-only, controls-only or combined case-control analyses. We call this conditional, induced correlation approach the Causal Pivot (CP) to indicate that the known cause is used to leverage the conclusion.

Simulation based power analyses enable controlled comparsion of the CP approach and alternatives. The CP applies in cases-only analyses; the Wilcox and Z-test are chosen as non-parametric comparators that also apply in cases-only analysis. The Wilcox test has been proposed previously [26,27], and focuses exclusively on the induced correlation between PRS and RV state conditional on disease. Alternatively, the Z-test focuses exclusively on the rate change in RV in the disease population compared to its expected frequency in the general population. Thus the Z-test ignores or fails to incorproate the PRS information. The CP approach incorporates both the frequency information from the general population and also the PRS-RV

conditional dependency. The results show that in the absence of a strong interaction the CP-LRT has stronger power than the Wilcox test, but it is weaker than the Z-test. When there is an interaction effect, the power of the CP-LRT exceeds the Z-test, but the Wilcox test surpasses it. Importantly, the CP likelihood approach results in parameter estimates and confidence intervals (see Supplemental materials); these interval estimates are an important advantage of the CP procedure over the non-parametric methods. Results show that the parameter estimates from the CP MLE in the logistic model framework are almost exactly equivalent to results from case-control logistic regression (see Supplmental Figure 2). Importantly, all of the tests have good power in the effect size range of interest for strong RV – where the RV effect exceeds that of the PRS.

The robustness analyses in Figure 2c and 2d are perhaps most revealing about the properties of the CP method. In the cases-only context, prior information concerning RV frequency is a precondition to apply either the Z-test or the CP-LRT. Importantly, the CP-LRT is informed by the structural model and includes the PRS as part of its procedure, while the Z-test does not. In terms of specificity, Figure 2c shows that CP-LRT is more stable against false positives when the population rate of RV is underestimated. Conversely, the Figure 2d shows that CP-LRT is also more sensitive when the RV rate is overestimated. Taken together Figure 2c and 2d show a key advantage of the CP-LRT. Because CP procedure is drawing information from the PRS it is effective when the Z-test fails.

The causal pivot is not a foreign concept in medical research.[6] Differential diagnosis is an example of causal pivot thinking. Given that an individual is observed to have some phenotype, when certain causes are ruled out, other causes become more likely. In genomic medicine, the causal pivot has been observed in somatic cancer, where driver mutations are

observed to exclude each other.[35-37] This exclusion principle is also at the heart of complementation group analysis in experimental genetics.

We chose polygenic risk scores (PRS)[38-43] as our pivot variable. PRS are quantitative metrics for individual subjects derived from the marginal associations of higher frequency markers with disease. To determine a PRS the individual's allelic counts are multiplied by weights derived fromby genome wide association studies (GWAS) – these weights themselves are commonly corrected for population structure and other covariates. PRS quantitatively represents the aggregate genetic risk from a large number of GWAS signals - summing a large number of diffuse or small effect size contributions. No single marker drives the PRS. Rather the PRS aggregates the additive genetic effects by including tens, hundreds, thousands or even millions of contributing variants. The advantage of the PRS is that it is encompassing and produces a straightforward causal interpretation. The disadvantage of the PRS is the disadvantage of any marker-based association analysis – the markers it employs are mostly common and have individually modest effects, while the effect of rare, private and *de-novo* variation is discounted or excluded.

RV are predominantly newer variants in a population and can have large functional effects. Studies of rare and Mendelian diseases are rooted in the large effects of rare variation as drivers of disease. RV are also observed in complex disease, where familial forms of many complex diseases – including the example diseases of Hypercholesterolemia (HC), Breast Cancer (BC)and Parkinson's Disease (PD). These diseases have familial and sporadic forms caused by large-effect RV.

Our simple three node causal pivot model makes a strong assumption that there are no common causes of the pivot variable and the causal candidate; in this case meaning that there are not common causes of the the PRS value and the RV state. We demonstrated that the assumptions can be relaxed and the causal pivot method can be extended to cases where the independence assumption may not hold. We considered global ancestry as an obvious

possible confounder illustrated in the ‘diamond’ graph (Figure 1B). The relationship between ancestry and the PRS may be complex and non-linear; the RV-ancestry relationship can be equally cryptic; as such linear approaches to control for the relationship may fail. We used the principal component vectors from genome wide ancestry analysis to implement non-parametric approaches to control for the possible role of ancestry confounding. We used both matching and inverse probability weighting where the procedure was determined by machine learning from the ancestry principle components. In both cases our results indiciated a highly significant induced negative correlation between PRS and RV after controlling for ancestry in the demonstration diseases (**Figure 4).**

The Causal Pivot approach extends to the context of non-binary candidate causes. We examined its use in the context of a pathway burden effect using the lysosomal storage pathway in PD. Extensive literature has demonstrated the importance of the lysosome in neurodegenerative disease, and many examples of Mendelian PD involve variants in lysosomal storage genes. We modeled the count of ClinVar pathogenic variants in the pathway as a count variable and we show that rate of the rare pathogenic variation negatively covaries with PRS among affected individuals. We also present a graphical view of the PRS as a function of the variant count and see that its mean behavior is decreasing in the variant burden.

We used simulation and real data to consider CP analyses in a cases-only context. The cases-only test has slightly less power than a comparator case-control analysis, indicating that most of the information about the causal effect of the RV is carried in the cases. It is important to note that such a cases-only analysis is not possible using traditional forward association tests. The new strategy is therefore ideal when selection on the phenotype is unavoidable. There are many such scenarios relevant to complex diseases. Genotype X environment interaction studies investigate whether the genetic variant and environmental exposure jointly influence the risk, which cannot be tested in those who do not have the outcome. Case-crossover designs the study is restricted to individuals who have already experienced the

outcome, and the focus is on understanding the exposure timing in relation to the event within the same individual. Cases only analysis also applies in the context of somatic mutation outcomes in cancer where analysis is confined to individuals who already have cancer, and the focus is on the correlation between mutations and exposures or germline genetic variation. Another cases only context occurs in drug safety studies that investigate the association between drug exposure specific adverse outcomes; the design also applies to pharmacogenetic studies that assess the drug response among patients who have already experienced a particular outcome, such as an adverse drug reaction. In all these circumstances cases-only genetic analyses are required.

The study of causal heterogeneity in complex disease genetics is a fertile area for research. Significant progress has been made in identifying genes and mechanistic processes that drive complex disease. In this context comparably less effort has been assigned to deconvolving the disease population into subgroups. Our CP and CP-LRT method exploits prior research from GWAS as well as surveys of genetic variation to drive discovery and assignment of causal types. Additional research in the area of causal heterogeneity using SCM is likely to prove productive.

Appendices

Derivation of Causal Pivot - .pdf

Supplementary Figures - .pdf

Table of Annotated Variants - .xlsx

Declarations of Interests

JWB and CAS are co-owners of Texas Genomics Consulting ~~which to date has worked in the development and implementation of clinical genetic testing in rare disease and cancer~~.

Acknowledgments

Grant from Genetics & Genomics Services, Inc.

Author Contributions

CAS developed the concept of the Causal Pivot and derived the SCM representation. Working together with TTT he developed the likelihood representation and likelihood ratio test. CAS derived the likelihood score equations for the likelihood analysis as well as the ratio representation for simplifying the CP-LRT; he developed the approach to Fisher Information and interval calculation for the method. He also developed the matching and inverse probability weighting method for Ancestry correction as well as the pathway burden modeling. He supervised the implementation of software and he wrote the manuscript and supplement. *Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Software, Resources, Supervision, Writing – original draft, Writing – review & editing*

CJW participated heavily in data curation as well as refinement and execution in software implementation of both power simulations and UKBB validations. CJW also executed the implementation of matching and random forest analysis to control for ancestry. He generated the results and figures for the manuscript. *Data curation, Software, Validation, Visualization, Writing – original draft*

JWB worked together with CAS to develop the concept of CP for complex disease and its context in genetic epidemiology. He guided the choice of example diseases and the selection of the proof of concept genes. He contributed to the derivation of the log-likelihood score equations. He provided funding support, and he was strongly involved in writing the manuscript. *Conceptualization, Data curation, Funding acquisition, Project administration, Software, Resources, Writing – original draft, Writing – review & editing*

JS contributed to development of the CP concept in neurodegenerative disease and identified *GBA* as a candidate gene for CP analysis in PD. He reviewed and assisted with the manuscript and provided funding support. *Conceptualization, Data curation, Funding acquisition, Writing – review & editing*

TTT worked with CAS on development of the LRT for CP and on initial simulation studies. He participated in development of the manuscript. *Formal analysis, Methodology , Writing – review & editing*

ND assisted with implementation of the log-likelihood score equations for both the logistic and liability model. He performed simulation studies for power analysis. *Formal analysis, Methodology , Writing – review & editing*

Web Resources

N/A

Data and code availability

Github ....

References

1. Selvaraj, M.S., Li, X., Li, Z., Pampana, A., Zhang, D.Y., Park, J., Aslibekyan, S., Bis, J.C., Brody, J.A., Cade, B.E., et al. (2022). Whole genome sequence analysis of blood lipid levels in >66,000 individuals. Nat Commun *13*, 5995. 10.1038/s41467-022-33510-7.
2. Goodrich, J.K., Singer-Berk, M., Son, R., Sveden, A., Wood, J., England, E., Cole, J.B., Weisburd, B., Watts, N., Caulkins, L., et al. (2021). Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. Nat Commun *12*, 3505. 10.1038/s41467-021-23556-4.
3. Medeiros, A.M., and Bourbon, M. (2021). Polygenic contribution for familial hypercholesterolemia (FH). Curr Opin Lipidol *32*, 392-395. 10.1097/MOL.0000000000000787.
4. Sun, N., and Zhao, H. (2020). Statistical Methods in Genome-Wide Association Studies. Annual Review of Biomedical Data Science *3*, 265-288. 10.1146/annurev-biodatasci-030320-041026.
5. Chen, W., Coombes, B.J., and Larson, N.B. (2022). Recent advances and challenges of rare variant association analysis in the biobank sequencing era. Front Genet *13*, 1014947. 10.3389/fgene.2022.1014947.
6. Pierce, B.L., and Ahsan, H. (2010). Case-only genome-wide interaction study of disease risk, prognosis and treatment. Genet Epidemiol *34*, 7-15. 10.1002/gepi.20427.
7. Wu, L., Schaid, D.J., Sicotte, H., Wieben, E.D., Li, H., and Petersen, G.M. (2015). Case-only exome sequencing and complex disease susceptibility gene discovery: study design considerations. J Med Genet *52*, 10-16. 10.1136/jmedgenet-2014-102697.
8. Webster, A.J. (2022). Causal attribution fractions, and the attribution of smoking and BMI to the landscape of disease incidence in UK Biobank. Sci Rep *12*, 19678. 10.1038/s41598-022-23877-4.
9. Glass, T.A., Goodman, S.N., Hernan, M.A., and Samet, J.M. (2013). Causal inference in public health. Annu Rev Public Health *34*, 61-75. 10.1146/annurev-publhealth-031811-124606.
10. Rothman, K.J., and Greenland, S. (2005). Causation and causal inference in epidemiology. Am J Public Health *95 Suppl 1*, S144-150. 10.2105/AJPH.2004.059204.
11. Smith, G.D., and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol *33*, 30-42. 10.1093/ije/dyh132.
12. Zheng, J., Baird, D., Borges, M.C., Bowden, J., Hemani, G., Haycock, P., Evans, D.M., and Smith, G.D. (2017). Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep *4*, 330-345. 10.1007/s40471-017-0128-6.
13. Sanderson, E., Glymour, M.M., Holmes, M.V., Kang, H., Morrison, J., Munafo, M.R., Palmer, T., Schooling, C.M., Wallace, C., Zhao, Q., and Smith, G.D. (2022). Mendelian randomization. Nat Rev Methods Primers *2*. 10.1038/s43586-021-00092-5.

14. Hu, X., Cai, M., Xiao, J., Wan, X., Wang, Z., Zhao, H., and Yang, C. (2024). Benchmarking Mendelian randomization methods for causal inference using genome-wide association study summary statistics. Am J Hum Genet *111*, 1717-1735. 10.1016/j.ajhg.2024.06.016.
15. Dudbridge, F. (2021). Polygenic Mendelian Randomization. Cold Spring Harb Perspect Med *11*. 10.1101/cshperspect.a039586.
16. Pingault, J.B., Richmond, R., and Davey Smith, G. (2022). Causal Inference with Genetic Data: Past, Present, and Future. Cold Spring Harb Perspect Med *12*. 10.1101/cshperspect.a041271.
17. Pearl, J. (2010). An introduction to causal inference. Int J Biostat *6*, Article 7. 10.2202/1557-4679.1203.
18. Petersen, M.L. (2011). Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. Epidemiology *22*, 378-381. 10.1097/EDE.0b013e3182126127.
19. Lipsky, A.M., and Greenland, S. (2022). Causal Directed Acyclic Graphs. JAMA *327*, 1083-1084. 10.1001/jama.2022.1816.
20. Holmberg, M.J., and Andersen, L.W. (2022). Collider Bias. JAMA *327*, 1282-1283. 10.1001/jama.2022.1820.
21. Cai, S., Hartley, A., Mahmoud, O., Tilling, K., and Dudbridge, F. (2022). Adjusting for collider bias in genetic association studies using instrumental variable methods. Genet Epidemiol *46*, 303-316. 10.1002/gepi.22455.
22. Coscia, C., Gill, D., Benitez, R., Perez, T., Malats, N., and Burgess, S. (2022). Avoiding collider bias in Mendelian randomization when performing stratified analyses. Eur J Epidemiol *37*, 671-682. 10.1007/s10654-022-00879-0.
23. Mahmoud, O., Dudbridge, F., Davey Smith, G., Munafo, M., and Tilling, K. (2022). A robust method for collider bias correction in conditional genome-wide association studies. Nat Commun *13*, 619. 10.1038/s41467-022-28119-9.
24. Mitchell, R.E., Hartley, A.E., Walker, V.M., Gkatzionis, A., Yarmolinsky, J., Bell, J.A., Chong, A.H.W., Paternoster, L., Tilling, K., and Smith, G.D. (2023). Strategies to investigate and mitigate collider bias in genetic and Mendelian randomisation studies of disease progression. PLoS Genet *19*, e1010596. 10.1371/journal.pgen.1010596.
25. Swanson, S.A., Robins, J.M., Miller, M., and Hernan, M.A. (2015). Selecting on treatment: a pervasive form of bias in instrumental variable analyses. Am J Epidemiol *181*, 191-197. 10.1093/aje/kwu284.
26. Zhou, D., Yu, D., Scharf, J.M., Mathews, C.A., McGrath, L., Cook, E., Lee, S.H., Davis, L.K., and Gamazon, E.R. (2021). Contextualizing genetic risk score for disease screening and rare variant discovery. Nat Commun *12*, 4418. 10.1038/s41467-021-24387-z.
27. Lu, T., Forgetta, V., Richards, J.B., and Greenwood, C.M.T. (2022). Polygenic risk score as a possible tool for identifying familial monogenic causes of complex diseases. Genet Med *24*, 1545-1555. 10.1016/j.gim.2022.03.022.
28. Spirtes, P., Glymour, C., and Scheines, R. (2001). Causation, Prediction, and Search, second edition (The MIT Press).

29. Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H., and Bühlmann, P. (2012). Causal Inference Using Graphical Models with the R Package pcalg. Journal of Statistical Software *47*, 1 - 26. 10.18637/jss.v047.i11.
30. Badsha, M.B., Martin, E.A., and Fu, A.Q. (2021). MRPC: An R Package for Inference of Causal Graphs. Front Genet *12*, 651812. 10.3389/fgene.2021.651812.
31. Thompson, D.J., Wells, D., Selzam, S., Peneva, I., Moore, R., Sharp, K., Tarran, W.A., Beard, E.J., Riveros-Mckay, F., Palmer, D., et al. (2022). UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. medRxiv, 2022.2006.2016.22276246. 10.1101/2022.06.16.22276246.
32. Madsen, A.M., Ottman, R., and Hodge, S.E. (2011). Causal models for investigating complex genetic disease: II. what causal models can tell us about penetrance for additive, heterogeneity, and multiplicative two-locus models. Hum Hered *72*, 63-72. 10.1159/000330780.
33. Stern, C. (1973). Principles of Human Genetics, 3rd Edition (W. H. Freeman and Company).
34. Woodward, A.A., Urbanowicz, R.J., Naj, A.C., and Moore, J.H. (2022). Genetic heterogeneity: Challenges, impacts, and methods through an associative lens. Genet Epidemiol *46*, 555-571. 10.1002/gepi.22497.
35. Babur, O., Gonen, M., Aksoy, B.A., Schultz, N., Ciriello, G., Sander, C., and Demir, E. (2015). Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biol *16*, 45. 10.1186/s13059-015-0612-6.
36. Dao, P., Kim, Y.A., Wojtowicz, D., Madan, S., Sharan, R., and Przytycka, T.M. (2017). BeWith: A Between-Within method to discover relationships between cancer modules via integrated analysis of mutual exclusivity, co-occurrence and functional interactions. PLoS Comput Biol *13*, e1005695. 10.1371/journal.pcbi.1005695.
37. Zhang, W., Zeng, Y., Wang, L., Liu, Y., and Cheng, Y.N. (2020). An Effective Graph Clustering Method to Identify Cancer Driver Modules. Front Bioeng Biotechnol *8*, 271. 10.3389/fbioe.2020.00271.
38. Abraham, G., Rutten-Jacobs, L., and Inouye, M. (2021). Risk Prediction Using Polygenic Risk Scores for Prevention of Stroke and Other Cardiovascular Diseases. Stroke *52*, 2983-2991. 10.1161/STROKEAHA.120.032619.
39. Lakeman, I.M.M., Rodriguez-Girondo, M.D.M., Lee, A., Celosse, N., Braspenning, M.E., van Engelen, K., van de Beek, I., van der Hout, A.H., Gomez Garcia, E.B., Mensenkamp, A.R., et al. (2023). Clinical applicability of the Polygenic Risk Score for breast cancer risk prediction in familial cases. J Med Genet *60*, 327-336. 10.1136/jmg-2022-108502.
40. O'Sullivan, J.W., Raghavan, S., Marquez-Luna, C., Luzum, J.A., Damrauer, S.M., Ashley, E.A., O'Donnell, C.J., Willer, C.J., Natarajan, P., American Heart Association Council on, G., et al. (2022). Polygenic Risk Scores for Cardiovascular Disease: A Scientific Statement From the American Heart Association. Circulation *146*, e93-e118. 10.1161/CIR.0000000000001077.

689 41. Rao, A.S., and Knowles, J.W. (2019). Polygenic risk scores in coronary artery
690 disease. Curr Opin Cardiol *34*, 435-440. 10.1097/HCO.0000000000000629.
691 42. Riveros-Mckay, F., Weale, M.E., Moore, R., Selzam, S., Krapohl, E., Sivley, R.M.,
692 Tarran, W.A., Sorensen, P., Lachapelle, A.S., Griffiths, J.A., et al. (2021). Integrated
693 Polygenic Tool Substantially Enhances Coronary Artery Disease Prediction. Circ
694 Genom Precis Med *14*, e003304. 10.1161/CIRCGEN.120.003304.
695 43. Roberts, E., Howell, S., and Evans, D.G. (2023). Polygenic risk scores and breast
696 cancer risk prediction. Breast *67*, 71-77. 10.1016/j.breast.2023.01.003.

Figure Titles and Legends

Figure 1 Core Logic of the Causal Pivot. **Panel a)** depicts the basic model with PRS as a known causal factor indicated by the solid line and arrow terminating at Y; RV is an independent factor whose effect on Y is unknown and being evaluated. **Panel b)** depicts the impact of the solid arrow from PRS to Y; the solid line indicates a correlative relationship between PRS and Y which manifests as a higher PRS in cases where Y=1 than in controls where Y=0. **Panel c)** depicts the model of unconditional independence between PRS and RV. The unconditional independence results in a lack of pattern between PRS and RV state, which can also be viewed as a lack of correlation in the rate of the RV+ as a function of the tertiles of the quantitative PRS. In **panel d)** we depict the circumstance where RV influences phenotype as indicated by the solid arrow from RV to Y; the red box around Y indicates conditioning on the phenotypic state. In the circumstance where RV drives Y and where the value of Y is conditioned upon (red box), then an induced correlation appears between PRS and RV. This collider-induced correlation is depicted by the dashed red line connecting PRS and RV. In **panel e)** we depict the rate change in RV+ status according to the value of Y as would result from a solid arrow connecting RV and Y. In **panel f)** we depict the collider-induced correlation between PRS and RV conditional on Y; this correlation manifests as a Y-conditional shift in PRS between RV+ and RV-; alternatively, the rate of the RV+ state covaries with PRS conditional on Y. This collider induced correlation between PRS and RV juxtaposes with the unconditional independence of the PRS and RV as shown in **panel c)**. The Causal Pivot recruits the collider induced correlation together with the rate change in the RV – as shown in **panel d --** as signal to infer the RV-Y relationship.

Figure 2. Simulation Analyses The figure presents simulation analyses comparing the CP-LRT to alternatives. Monte Carlo methods are used to generate the results in panels a-d; these simulations generate PRS and RV status which are then used to generate trait outcomes under

the corresponding probabilistic model – either logisitic or quantitative. Alternative analyses are then performed to detect the effect of the RV factor and to compare alternative approaches on the same simulated data. Detailed markdown documentation and code to produce these analyses are provided (https://github.com/chadashaw/pivotr). Panel a) focuses on binary traits; the outcome is generated by a logistic model where PRS and RV are causal factors. Total population size is n=500,000, comparable to a UKBB scale. We set the intercept α=-2.2 corresponding to a disease prevalence of 10%, and the PRS effect $\beta=0.5$, consistent with UKBB data for our three demonstration diseases; the RV state is generated with frequency ω=0.001. The x-axis represents the main effect of the rare variant -- parameterized as γ. The y-axis represents the statistical power as determined by the proportion of simulation runs that reject the null hypothesis of no PRS effect under a nominal Type 1 error rate of 0.05. The curves present power comparisons to detect the RV effect as the main effect of the RV varies. The cases-only CP-LRT is shown in red. Cases-only alternatives are the Wilcox test (purple) that considers non-parametric mean shift in PRS values in cases between RV+ vs RV-; the Z-test (cyan) considers the increase in the frequency of RV+ among cases compared to what would be expected from the background frequency ω=0.001. Also included is the forward logistic regression using both cases and controls (green) that represents a typical regression approach. For the dashed line curves the PRS-RV interaction is set to η= -.4. In the absence of an interaction effect the Wilcox test has sharply lower power than CP-LRT and the Z-test. The Z-test is more sensitive to detect RV effects than the CP-LRT; the Z-test is also non-parametric and does not distinguish main effects and interaction effects between PRS and RV state. The Panel b) presents analyses when the phenotypic outcome is quantitative; again, a biobank scale

population is simulated, but here the trait is generated according to a linear model; we set $\alpha=0$ and set $\beta=0.3$; RV occurs at a rate of 0.001. To construct a conditional outcome we identify individuals who have a trait value larger than a quantile cutoff, and we use an empirical percentile of 0.93 consistent with the HC cutoff from UKBB corresponding to LDL> 190. We compare CP-LRT to cases-only alternatives of the Wilcox test (purple) and a Z-test (cyan) as well as an unconditional linear regression model (green) for the trait using the full sample. The CP-LRT has superior power to the Wilcox test in the absence of an interaction term. The Z-test still outperforms; the data linear regression has the best power. When the interaction effect is present all tests have higher power. The Wilcox test is notably higher. In panel c) we scrutinize fold increase in Type-1 error (y-axis, ratio of simulated type 1 error to 0.05) as a measure of the comparative stability of the CP-LRT compared to the Z-test. In c) simulations use the same parameters α, β,ω as in panel a) but holding γ,η at 0. Both the CP-LRT and Z-test require specification of ω the population rate of RV. The x-axis in c) considers ω^* where the analysis underspecifies the RV rate. This under specification results in a sharp increase in the type 1 error of the Z-test compared to the CP-LRT; the Z-test error immediately exceeds 0.05 as ω^* is decreased from the true value of 0.001. Differently, the CP-LRT – which leverages the collider induced correlation – is more robust and retains a type 1 error of close to 0.05 beyond a 10% reduction in ω. The red curve is uniformly below the Z-test indicating the CP-LRT retains better type 1 error robustness. In **panel d)** we again focus on the logistic model as in **panel a).** Here we compare the sensitivity of CP-LRT to Z-test when the RV has an effect but where the analysis false misspecifies the RV rate ω^* (x-axis) above its true value of 0.001 (dahsed vertical line). Incorrect overspecification of RV rate diminishes the power of the Z-test. In these simulations

we again set $\alpha=-2.2$; we set $\beta=0.6$ and $\gamma=2.2$, consistent with the parameterization for the *BRCA1* gene in BC (Supplemental Table 1). The vertical dashed line near $\omega^*=0.045$ indicates performance when the RV rate is set to the expected frequency of the RV in the Y=1 population. In this case the rejection rate of Z-test is about 0.05, while the CP-LRT has about 75% power. The retains CP-LRT power because it recruits information from the istructurally nduced correlation between RV and PRS.

Figure 3. **Demonstration of the Causal Pivot in UKBB data.** Panel a) shows a 3x3 collection of barplots where the vertical bars represent the count of RV+ Samples within each PRS-tertile grouping which defines the X-axes. Each row represents a different disease cohort paired with its disease specific PRS; HC in row one, BC in row two and PD in row three. Each column presents a different disease gene as indicated by the column headers, with *LDLR* variation tabulated in column one, *BRCA1* variation in column two and *GBA* in column three. The diagnonal of panel a) shows the negative correlation between the number of RV+ samples and PRS scores within each PRS-Disease-Gene grouping and a lack of association characterizes the barplots off the diagonal; the CP-LRT was used to determine p-values for all 9 analyses, and the p-values for all tests are p<10^-5 along the diagonal. Off diagonal entries are not significant except for the gene *LDLR* which shows modest association pattern with both BC and PD with $p<0.05$. More detailed results including confidence intervals for γ and η appear in the Supplemental materials. In panel cb the graphic presents the multiplicity and typology of the RV represented in the panel a) data. Most RV occur in only a single sample. A variety of different variant types are present. A detailed listing of the variants that inform the analysis is provided in a Supplemental Variant Table.

Figure 4 Controlling for Ancestry In any genetic association analysis ancestry is a potential common cause. In Panel a) we depict the scenario where ancestry represented by node A drives both PRS and RV state. In panel b) RV does not drive Y, and conditioning on both A and Y there is no induced or residual correlation between PRS and RV. In panel c) RV drives Y, and conditioning on both A and Y yields an induced correlation between PRS and RV. In panel d) we present three distinct rows of analyses; HC in row one, BC in row two and PD in row three. Each row shows three results. The scatter plots show RV^- (black points) and RV^+ (red points) cases as points in an ancestry visualization space. The scatter plot representation is determined by the first 5 PCs reduced to two dimensions using t-distributed stochastic neighbor embedding (t-SNE). A zoom in view of each row is indicated by the excerpted rectangular region. Each RV^+ case is matched to the nearest (k=2) RV^- cases in the 5 dimensional ancestry space. The mean difference in the PRS of each RV+ and its ancestry matched neighbors is determined. The average of these PRS differences is then calculated and shown in the dashed vertical red line in each histrogram plot. The histogram plots show both these real data calculations as well as the results of permutation testing depicted in the histrograms. The permutation distributions are generated by randomly permuting the RV states of each sample in the cohort while fixing the ancestry data; the procedure of identifying neighbors and calculing the mean difference in PRS is repeated. In addition to the histograms that present the permutation distributions, the 5% threshold value is shown by the dashed black lines. The observed value from the true RV+ ancestry matched data is shown by the dashed red line is significantly below the 0.05 quantile in all cases, consistent with the pattern that RV+ cases have lower PRS than their RV- ancestry-matched neighbors.

Figure 5. Pathway burden analysis. Panel a) shows the the PRS values (y-axis) against the count of pathogenic/likely pathogenic RV in the lysosomal storage pathway among PD+ cases from UKBB. Each boxplot presents cases with the tabulated count of RV+ per case. There is a decreasing trend in PRS, and a Poisson regression analysis which models the rate of RV as a function of PRS demonstrates a statisticallly significant negative association between the rate of RV+ and the PRS value (p=0.03). Panel b) depicts the sparse occupancy of positivity for 2 variants. No single pair accounts for the majority of instances.

Tables

Rare Variant Table from UKBB

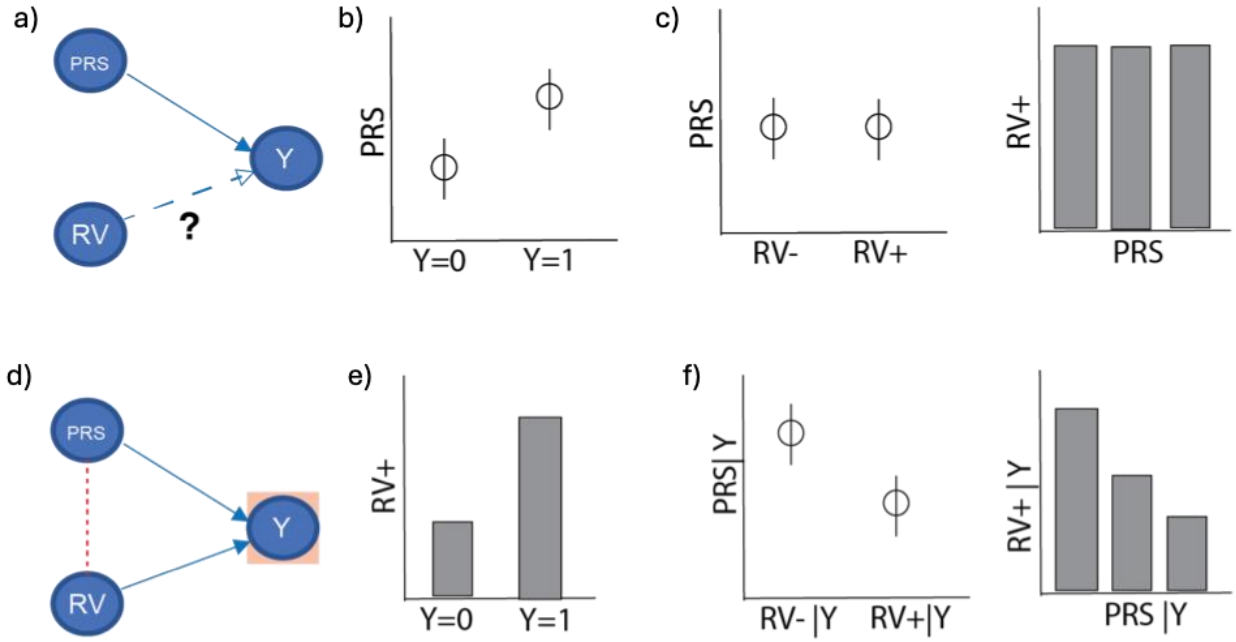

Figure 1: Core Logic of the Causal Pivot

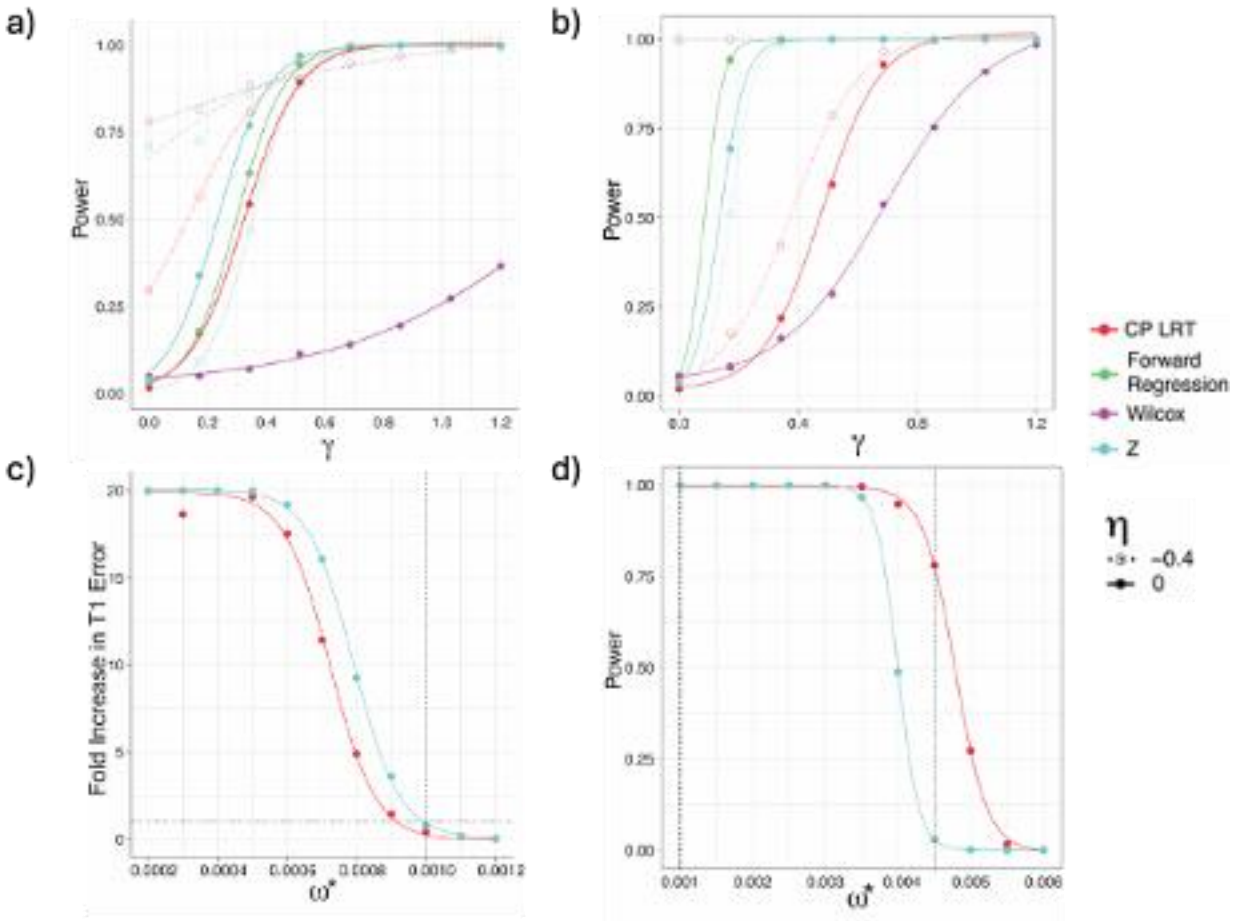

Figure 2: Simulation Analyses

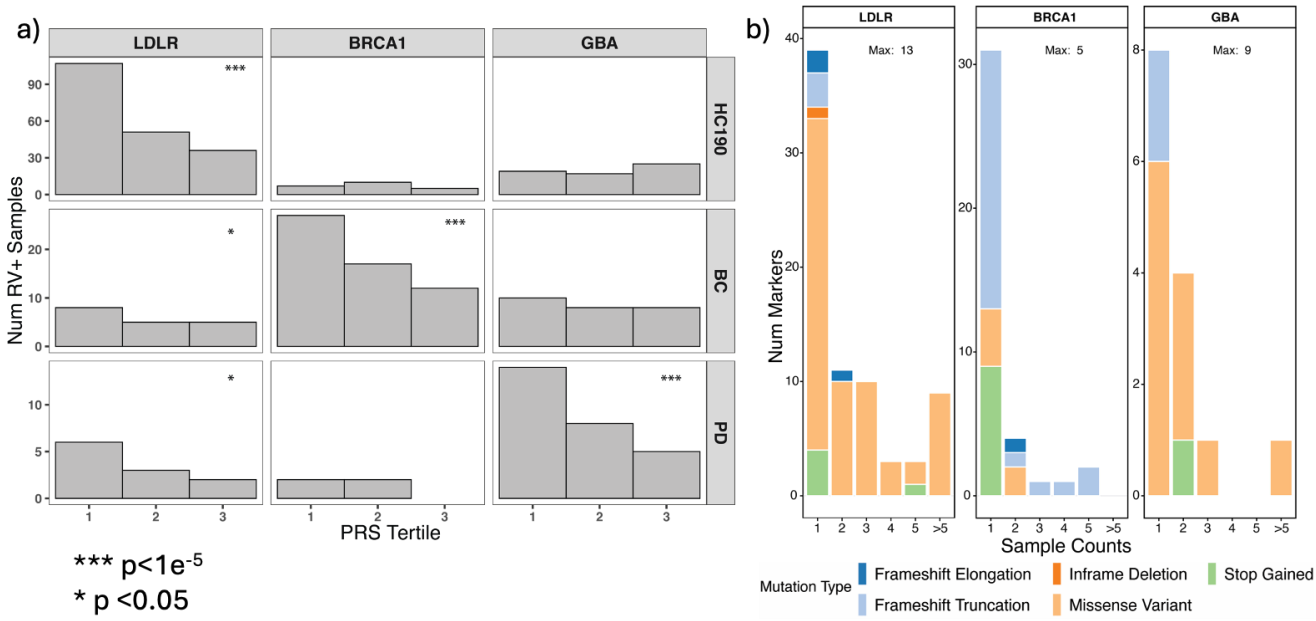

Figure 3: UKBB Results

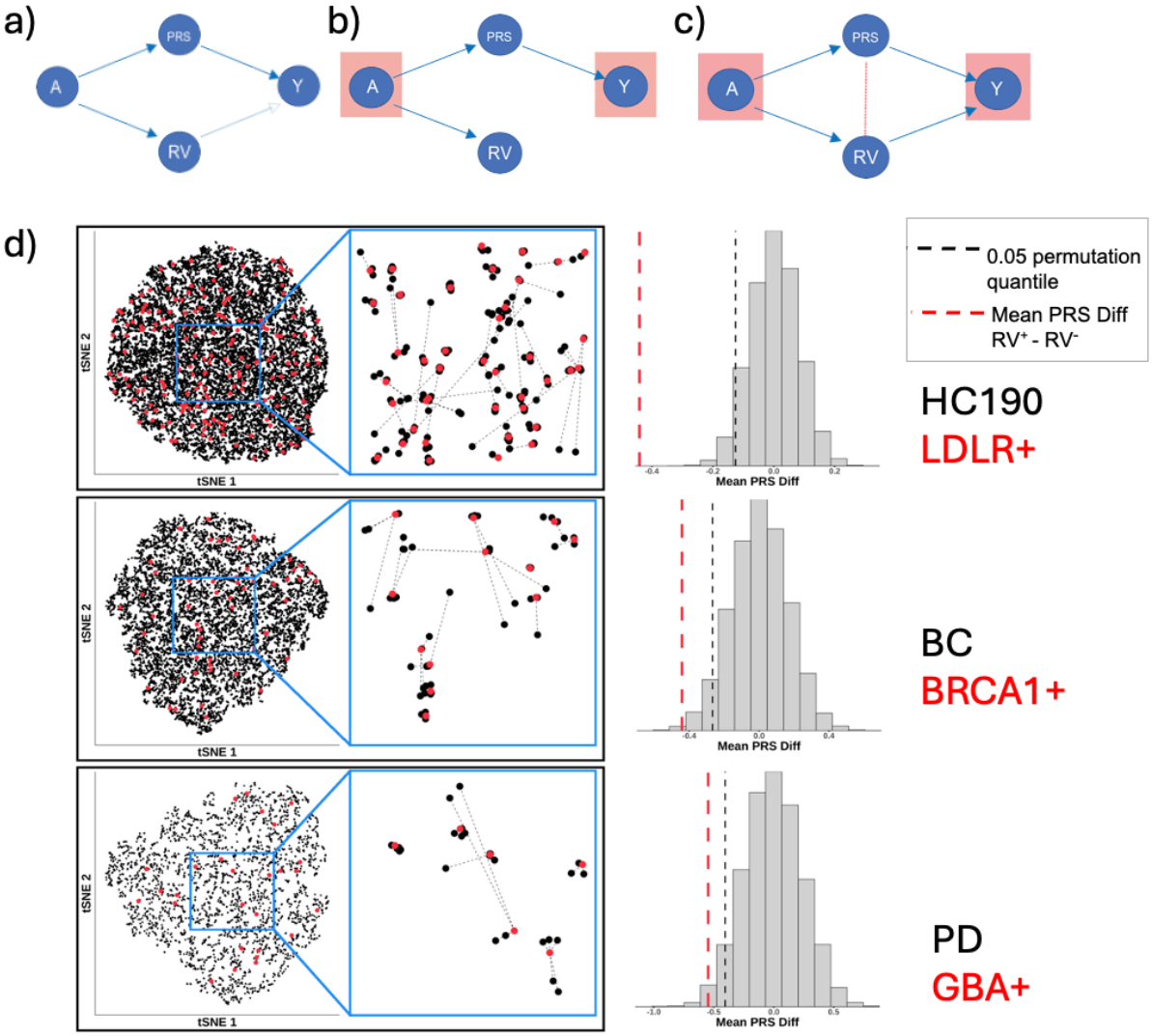

Figure 4: Controlling for Ancestry

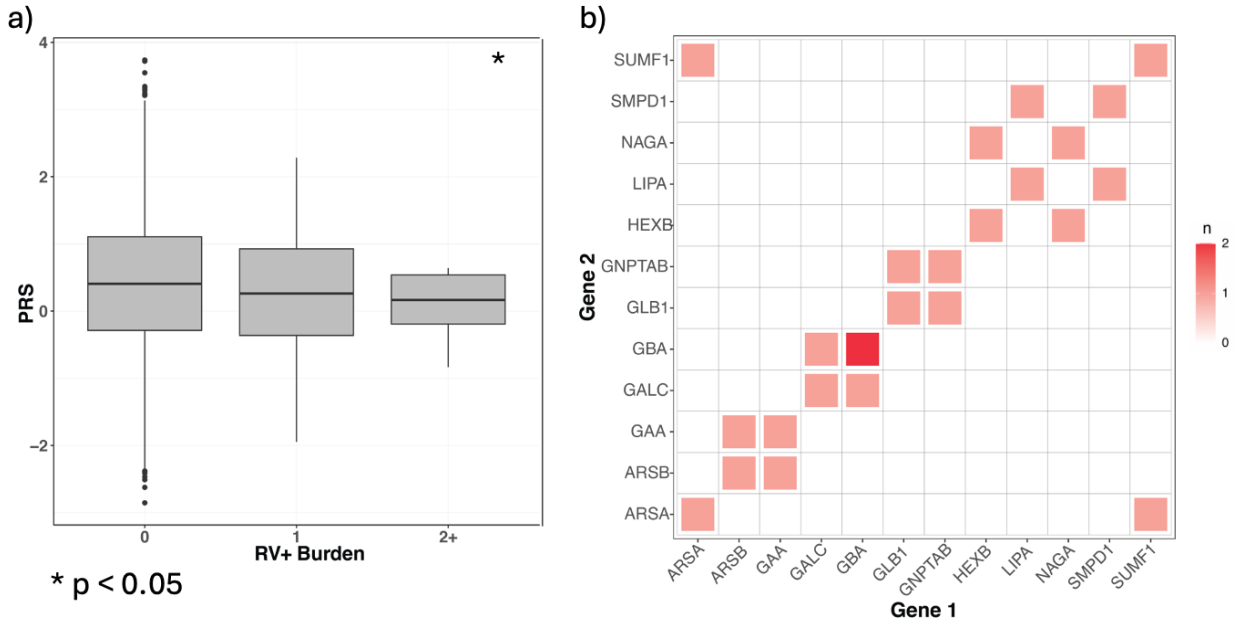

Figure 5: Pathway Burden