Received: 9 May 2023 | Revised: 30 December 2023 | Accepted: 27 March 2024

DOI: 10.1002/gepi.22563

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

A novel application of data-consistent inversion to overcome spurious inference in genome-wide association studies

Negar Janani[1] | **Kendra A. Young**[2] | **Greg Kinney**[2] | **Matthew Strand**[3] | **John E. Hokanson**[2] | **Yaning Liu**[1] | **Troy Butler**[1] | **Erin Austin**[1]

[1]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, USA

[2]Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado, USA

[3]Division of Biostatistics, National Jewish Health, Denver, Colorado, USA

Correspondence
Negar Janani, Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, USA.
Email: negar.janani@ucdenver.edu

Funding information
COPDGene Project; NHLBI, Grant/Award Numbers: U01 HL089897, U01 HL089856; COPDGene, Grant/Award Number: NCT00608764; National Science Foundation, Grant/Award Number: DMS-2208460

Abstract

The genome-wide association studies (GWAS) typically use linear or logistic regression models to identify associations between phenotypes (traits) and genotypes (genetic variants) of interest. However, the use of regression with the additive assumption has potential limitations. First, the normality assumption of residuals is the one that is rarely seen in practice, and deviation from normality increases the Type-I error rate. Second, building a model based on such an assumption ignores genetic structures, like, dominant, recessive, and protective-risk cases. Ignoring genetic variants may result in spurious conclusions about the associations between a variant and a trait. We propose an assumption-free model built upon data-consistent inversion (DCI), which is a recently developed measure-theoretic framework utilized for uncertainty quantification. This proposed DCI-derived model builds a nonparametric distribution on model inputs that propagates to the distribution of observed data without the required normality assumption of residuals in the regression model. This characteristic enables the proposed DCI-derived model to cover all genetic variants without emphasizing on additivity of the classic-GWAS model. Simulations and a replication GWAS with data from the COPDGene demonstrate the ability of this model to control the Type-I error rate at least as well as the classic-GWAS (additive linear model) approach while having similar or greater power to discover variants in different genetic modes of transmission.

KEYWORDS

COPD, data-consistent inversion, genome-wide association study, inverse problem, uncertainty quantification

1 | INTRODUCTION

We develop and establish the credibility of a novel method for genome-wide association studies (GWAS). The usual GWAS utilizes regression, requiring multiple assumptions, such as normality of errors and specifying the genetic model between single-nucleotide polymorphisms (SNPs) and phenotypes. However, these assumptions are sometimes not met, causing inflation in false positives and diminished power. We propose a substantial modification to the traditional GWAS. We use a similar linear model framework to preserve features like interpretability and ability to add

covariates, while removing many of the usual linear assumptions. However, instead of least squares estimation, we estimate the sampling distribution of mean effects related to SNPs using data-consistent inversion (DCI), a new method from a field of mathematics known as uncertainty quantification.

The following work is organized in distinct sections in an attempt to provide detailed discussion for each key element of the problem, but also allow those with sufficient background to feel confident in skipping a section. Section 2 reviews the history of GWAS along with technical details of the linear model it uses. This includes a description of the limitations of the usual GWAS to set up the context for how DCI can be used to overcome these. Section 3 provides an overview of uncertainty quantification before delineating DCI and the specifics of how we implement it. Section 4 is a rigorous simulation experiment comparing the classic GWAS to our DCI-based GWAS. Section 5 applies this approach to real data from the COPDGene Project where we replicate and extend a GWAS done in that setting. Section 6 contains a summarizing discussion.

2 | GENOME-WIDE ASSOCIATION STUDIES (GWAS)

Identifying associations between genetic variation and disease engenders opportunities to develop interventions to help those affected by the disease. GWAS is a regression-based approach used to find an association between SNPs, and phenotypes of interest. GWAS involves hypothesis testing of all available SNPs individually across the genome with one specific phenotype. Regression coefficients provide an estimate of the mean effect of the SNP's alleles on the phenotype; thus, coefficients significantly different than zero identify possible genotype–phenotype associations. The validity of these results depends on the appropriateness of the underlying linear model and assumptions.

2.1 | GWAS limitations

GWAS results have been constrained by the inherent limitations of linear regression. Moreover, flexibility regarding (or even removing) the linear model assumptions has the potential to enhance the ability of GWAS to discover new SNPs–phenotype associations. If the linear model uses a different genetic structure than the true underlying model of transmission, then the rates of spurious associations are possibly inflated (Ray & Chatterjee, 2020). Furthermore, the classic incorrectly treating variables (e.g., ordinal factors) as continuous increases bias and decreases power in GWAS (Verhulst & Neale, 2021). A consequence of ignoring dominance effects is a biased estimation and an increase in false-positive rates (Monir & Zhu, 2017). It is not feasible to individually specify the genetic model when analyzing millions of SNPs, thus, defaulting to the least problematic additive form is reasonable, but this will inevitably lead to false conclusions about the associations between nonadditive SNPs and a phenotype. Replacing the additive model with the full genetic model (Monir & Zhu, 2017) coupled with limiting the proportion of false-positive results among all positives to some small value by Bayesian GWAS (Fernando & Garrick, 2013) are suggested as ways to control false-positive results. Ideally, though, a method would not need any genetic model to be specified, no matter if a Bayesian or frequentist approach is used. Our adaptation is not affected by this underlying structure.

At its core, GWAS is a regression-based method that uses Ordinary Least Squares (OLS) to estimate the coefficients (the SNP's mean effect) of the linear model. For OLS to be best, linear, unbiased, and estimator, the Gauss–Markov theorem requires errors to be uncorrelated, have equal variances and mean of zero to guarantee the (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators. Furthermore, there is an implicit assumption that the residuals, the error between the observed and predicted phenotype values, follow a normal distribution. The validity of the t test for each SNP depends on the normality of residuals; otherwise, we cannot ensure the normality of the sampling distribution for our estimate of the SNP's mean effect ($\hat{\beta}_{SNP}$). On the basis of the residual assumptions, $\hat{\beta}_{SNP}$ is normally distributed with mean β (hence unbiased) and the standard deviation of $\sigma_{\hat{\beta}_{SNP}}$, that is,

$$\hat{\beta}_{SNP} \sim N\left(\beta_{SNP}, \sigma^2_{\hat{\beta}_{SNP}}\right). \quad (1)$$

Although the central limit theorem mitigates deviations from normality, it is an asymptotic result. Consequently, it may require a potentially unobtainable sample size to guarantee that the sampling distribution is close enough to normality for accurate conclusions at the genome-wide significance threshold ($\alpha = 5 \times 10^{-8}$). Even small violations of this innocuous normality assumption have the potential to severely and adversely affect both the Type-I error rate (false positives) and the power of the hypothesis test to correctly identify SNPs associated with a phenotype (increase the Type-II error rate) or false-negative. Phenotypes often have markedly non-normal distributions, and the central limit theorem approximation is unreliable for GWAS regression p values (Connor & O'Neill, 2017). To provide one example relevant to our simulation study: right-skewed and non-

normal $\hat{\beta}_{SNP}$ s increase the mean (β) and the estimate of the standard deviation (SD) becomes inflated by the small proportion of observations in the drawn-out upper tail. The resulting inflation in the estimated standard error (SE) leads to a deflated value for the t statistic, causing Type-II errors (Fayers, 2011). Although the regression-based t test has Type-I error that is robust to deviations from normality for reasonably sized samples, the robustness of the power of the t test in the presence of non-normality is not as well-guaranteed as the robustness of the Type-I error (Feingold, 2002). Our proposed method adapts the linear structure to the non-normality anywhere in the model.

Multiple statistical procedures exist to address the problem of non-normality. Transformation is a common method used to remedy the problem of non-normality of phenotypes. Another strategy is to perform a parametric transformation to approximate the normality of quantitative traits (Goh & Yap, 2009). Previous research, however, has demonstrated that transformation may not solve non-normality issues in the regression model (Pires & Rodrigues, 2007).

Other methods have attempted to handle the violations of normality and the incorrect specification of the genetic model. Robust regression methods have been successful in handling outliers, but performed poorly in real data settings where results were expected to be replicated (Lourenço et al., 2011). Generalized Estimating Equations are another natural choice, but they can be more appropriate when there is a correlation among the observations, for example, from a pedigree study (Murabito et al., 2007) or a longitudinal phenotype (Sitlani et al., 2015). To remove the dependency on these assumptions, nonparametric methods have been developed. One such method applies a Marcus-like contrast matrix to represent three modes of inheritance (dominant, recessive, and additive) where the contrasts allow testing of a nonparametric estimation of SNP–phenotype association for all three modes (Konietschke et al., 2012). The method is effective except that it must test for the three models, and is impacted by multiple testing constraints. Moreover, the method does not allow for adjustments for covariates. We retain the linear structure to, among other reasons, allow covariates. Lastly, newer methods increase power by pooling SNPs from a gene for a single phenotype such as the method of multivariate multiple linear regression (RMMLR) (Basu et al., 2013) or pooling suspected related traits for a single SNP, such as the scaled multiple-phenotype association test (SMAT) (Schifano et al., 2013). However, our goal is to maximize the power with respect to discovery in the GWAS single SNP setting. We also borrow from nonparametric methods, but without the requirement to test for multiple possible genetic structures.

On a technical note, linear regression has an additional assumption of equal variance of the random errors, which is also referred to in the literature as homoscedasticity. Stratification is one remedy if this assumption is violated (Behrens et al., 2011). Thus, we also employ stratification in our real data analysis, specifically in the case of heteroscedasticity across our two populations.

2.2 | In Section 3

We show how we remove these assumptions while maintaining the linear framework, which is desirable given its ease of interpretations (means) and ability to be easily applied genome-wide. Specifically, we build a computational mathematics-based GWAS that does not require the usual regression assumptions, which we now explain as background for our GWAS adaptation. The exact form of the SNP used in the linear regression varies depending on the investigator's assumptions about the way a genotype affects a trait. The genotypes (or minor allele counts) for an SNP can also be grouped into genetic classes or models, the most common being dominant, recessive, and additive models. Briefly, to describe these genetic models, assume **A** and **a** are the two possible alleles for an SNP (let **a** denote the minor, or less frequent, allele). A dominant model (for **A**) assumes that having one or more copies of the **A** allele increases the risk for a disease compared with **a** (i.e., **aA** or **AA** genotypes have approximately an equivalent higher risk than **aa**). The recessive model (for **A**) assumes that two copies of the **A** allele are required to alter risk. The additive model (for **A**) assumes that there is a uniform, linear increase in risk for each copy of the **A** allele. A common practice for GWAS is to assume additive model are correct for every SNP, as the additive model has reasonable power to detect both additive and dominant effects (Bush & Moore, 2012).

3 | UNCERTAINTY QUANTIFICATION AND DCI

As computational models are increasingly relied upon to describe biological systems and genetics, quantifying the uncertainties in model inputs that influence solution is critical. Inverse propagation problems use observable output data to infer the model input parameters that likely generated those data and calibrate the parameters for distributions related to those inputs. Uncertainties in observable data often necessitate the formulation of a *stochastic* inverse problem where the solution is a computationally simple methodology based on density estimation. In this work, we apply a recently developed measure-theoretic framework and its density-based solution methodology,

which is referred to as *DCI* to solve a stochastic inverse problem for GWAS. Formally, DCI constructs a probability measure on model input parameters whose predicted outcome matches (is consistent with) a probability measure on the observable outputs of the model. We summarize some details below and direct the interested reader to Butler et al. (2020; and the references therein) for more details related to a stochastic mapping version of DCI that is most pertinent to this work.

The main goal of DCI is to infer information about the parameters of a model from observations. DCI is a data-adaptive computational method to determine the distribution of possible model parameter values which reproduce a distribution of model outputs associated with a distribution of observed values. More rigorous details are provided shortly. To motivate the general idea, the process starts with a user specifying the model and initial distributions of the model parameters. A "predicted" distribution of model outputs is obtained by propagating the initial distribution through the model. DCI then reconciles the differences between the predicted and observed distributions to update the initial parameter distribution in a specific probabilistic way such that propagating this updated distribution back through the model reproduces the observed distribution associated with model outputs. This "reproduction" of a distribution of model outputs is what is referred to as the "consistency criteria" in DCI. In other words, if the specified distribution of model outputs comes from an associated distribution of observed data, then the DCI result is a distribution that is "data-consistent" in the sense that the model maps the result onto the distribution of observed data.

This recently developed measure-theoretic framework for solving a stochastic inverse problem has been used in models of material science to determine process–structure–property linkages, one of the key objectives in material science, and infer a distribution of acceptable (consistent) microstructures, which expands the range of feasible designs in a probabilistic manner (Tran & Wildey, 2021). It has been also used to describe and quantify the impact of measurement error on predictions of lung function in the COPDGene cohort (Zachary et al., 2022).

3.1 | Connection between GWAS model and DCI

As previously mentioned, linear or logistic regression models are utilized within GWAS to test for associations between variants and traits, depending on whether the phenotype is quantitative (such as height, blood pressure, or body mass index) or dichotomous (such as the presence or absence of a trait), respectively. Covariates such as age, sex, and ancestry are included to avoid confounding effects from these or other factors.

The classic-GWAS model utilized in this study is a linear regression model. Thus, it is considered as a solution to a type of stochastic inverse problem because the observed output data, the phenotype of interest, supervises input data to estimate the coefficient parameter of the SNP in the linear model. In this work, we use the same linear model for the DCI-based solution except that we only assume the linear form and nothing about the distribution of the errors or correctness of the additive genetic model assumption. Our contention is that we can retain the value of the linear form, but provide an alternative way to estimate its parameters that is not dependent on potentially false assumptions. Thus, our method is in essence replacing OLS with DCI in the classic-GWAS model, resulting in, we believe, more defensible and likely more accurate estimates of SNP effects.

3.2 | How DCI addresses the non-normality problem

To translate the regression-based GWAS into DCI notation, we emphasize a key change in that DCI assumes the SNP's coefficients may vary (i.e., are treated as random variables from a particular uncertain distribution). In the DCI notation, Q denotes a quantity of interest, which is an uncertainty quantification term usually defined as the mapping from model parameters to model outputs associated with observable data. The linear form in the classic-GWAS model is used in this study for Q. This is done because our goal is to enhance and not replace the usual GWAS approach. Next, assume Q takes parameter values in the parameter space, denoted by $\lambda(\beta_{SNP})$, and uncertain (but knowable) variables, x, as the model inputs ($\boldsymbol{X_{SNP}}$) and then maps them to the space of estimated outputs ($\hat{\boldsymbol{Y}}$). In summary, $Q(x, \lambda)$ is the map between the parameter space β_{SNP} and the output space $\hat{\boldsymbol{Y}}$ given values for $\mathbf{X_{SNP}}$. The five steps taken by DCI to formulate and solve the stochastic inverse problem, resulting in estimates of β_{SNP}, are given below.

Step 1: Make initial assumptions about the parameter distributions, which we denote $\pi_{init}(x, \lambda)$, of the GWAS model parameters, such as β_{SNP}, based on the background knowledge of the model. Because the traditional GWAS model is a regression-based model, we initially assume a normal distribution for the for parameter β_{SNP} that is approximately the asymptotic distribution of $\hat{\beta}_{SNP}$ from OLS. This highlights one of the

strengths of our DCI approach. Specifically, the initial assumptions are essentially those that would lead to the same conclusions as classic GWAS. That is, we start our method by assuming GWAS is correct; for example, there are no issues with normality. We then use DCI to change, or update, to other distributions if the actual (observed) data is not compatible (i.e., inconsistent) with these assumptions.

Step 2: Use the GWAS linear model (denoted by Q) to create a set of initial predictions for the observed data based on the parameter distribution assumptions made in the first step. Specifically, generate a large number of samples from the initial parameter distributions; for example, 10,000 samples are generated from our initial normal distribution for the parameter $\hat{\beta}_{SNP}$. Each $\hat{\beta}_{SNP}$ randomly drawn from the initial normal distribution defines a unique candidate model, leading to a set of predicted outputs $\hat{\mathbf{Y}}$ for our given values for x, which are predictions of the observed output ($\mathbf{Y}$).

Step 3: Construct density estimates of the observed data $\mathbf{Y}$, label it $\pi_{obs}(Q(x,\lambda))$, and the set of predictions of the observed data, $\hat{\mathbf{Y}}$, label it $\pi_{predict}(Q(x,\lambda))$. In this work, we utilize a standard Gaussian kernel density estimator (GKDE) to construct non-parametric distribution estimates of both these densities. In other words, Step 3 in this work is to use a GKDE to construct an estimate of the distributions that produced both the observed data and the "push-forward" of the samples drawn from the initial density (referred to as the predicted density) generated in Step 2.

Step 4: Construct a "discrepancy ratio" $r(x,\lambda)$ defined as the ratio between the observed and predicted densities found in the previous step that are both evaluated on the model. This ratio is used to rescale the likelihoods of the initial distributions of parameters and the predicted data.

$$r(x,\lambda) = \frac{\pi_{obs}(Q(x,\lambda))}{\pi_{predict}(Q(x,\lambda))} \tag{2}$$

Step 5: Update the initial distributions of parameters $\hat{\beta}_{SNP}$ and in turn the predicted data $\hat{Y}$.

This reweighting process improves the fit of the associated predicted distribution of $\hat{\mathbf{Y}}$ to the true distribution of observed outputs of $\mathbf{Y}$. During this process, parameter values associated with predicted quantities of interest $Q(x,\lambda)$, where $\pi_{predict}(Q(x,\lambda)) > \pi_{obs}(Q(x,\lambda))$ are functionally down-weighted in the updated distributions, although parameter values associated with $Q(x,\lambda)$, where $\pi_{predict}(Q(x,\lambda)) < \pi_{obs}(Q(x,\lambda))$ are up-weighted in the updated distributions. Through the updated distribution, we assign new likelihoods to β_{SNP} as the true parameter value; that is, we have functionally created a distribution for $\hat{\beta}_{SNP}$ (and will switch to this notation going forward). At a computational level, an updated density, π_{update}, is immediately obtained after constructing the ratio $r(x,\lambda)$ from Step 4. The updated density usually is written in the form

$$\pi_{update}(x,\lambda) = \pi_{init}(x,\lambda) r(x,\lambda). \tag{3}$$

Equation (3) is not merely a formal construction as it also allows us to utilize standard rejection sampling techniques to produce samples that follow this updated distribution as a subset of the samples drawn from the initial distribution. Specifically, we utilize the ratio r to perform rejection sampling on the parameter values $\hat{\beta}_{SNP}$ s. This occurs in the data space, that is, the rejection sampling is applied on the set of predicted outcomes, $\hat{\mathbf{Y}}$s, which ensures that we keep parameter samples associated with predictions that are from a distribution *consistent* with the observed outputs. See Butler et al. (2020) for more details.

In summary, DCI concentrates on the distribution of model inputs and parameters that generate exactly the distribution of observed outputs. It begins with plausible distributions of parameter values, and then reweights them based on their compatibility with the observed data, creating an updated distribution reflecting the relative consistency between a candidate parameter and the observed data. The updated density is the solution to the stochastic inverse problem (Butler et al., 2020). Here, we specifically use DCI to create an alternate distribution of $\hat{\beta}_{SNP}$, one that shows how inputs such as SNPs could lead to outputs such as phenotypes by tuning the model parameters to be consistent with what we know for sure, which is the data we possess. All of this is done without any of the assumptions needed for classic-GWAS inference despite the utilization of these assumptions to create plausible initial distributions on model parameters. Moreover, the advantage of DCI over methods using a Bayesian framework is to establish a distribution on model inputs that exactly propagates to an observed distribution on model outputs although the Bayesian methods focus on a single point estimate of model inputs instead of a distribution on model inputs that propagates to the distribution of observed data (Zachary et al., 2022).

3.3 | How DCI addresses the additive assumption

We use the cell means model for the linear structure. The cell means model does not fit an overall mean effect for each additional minor allele. Instead, we estimate an individual mean effect for each of the SNP levels defined by minor allele counts of 0, 1, or 2, in effect treating the minor allele count as a categorical variable. More specifically, we identify the distribution of mean effects within three groups, one for each SNP level. This imposes no assumption about the underlying genetic model, which is conceptually how we allow DCI to remove the additive assumption in the classic-GWAS model. Specific mathematical details are given below.

The mathematical model that describes the relationship between the observed outcome $\boldsymbol{Y}$ and the SNP vector $\boldsymbol{X}_{SNP}$ for the cell means model corresponding to the GWAS study with 0, 1, and 2 SNP levels is given by

$$\boldsymbol{Y_j} = \mu_j + \boldsymbol{\epsilon_j}, \quad \boldsymbol{\epsilon_j} \sim N\left(0, \sigma_j^2 \boldsymbol{I_j}\right). \tag{4}$$

Here, $\boldsymbol{Y_j}$ is the vector of observed outcome or traits corresponding to the SNP level $j \in \{0, 1, 2\}$ corresponding to the number of minor alleles; μ_j is the mean effect for all observations of SNP level j; and $\boldsymbol{\epsilon_j}$ is the vector of random error for SNP level j; σ_j^2 is the variance of observations of SNP level j; and $\boldsymbol{I_j}$ is an $n_j \times n_j$ identity matrix, where n_j the number of observations with SNP level j. We estimate the model for each of the three SNP levels individually. That is, we use DCI to estimate the distribution of $\hat{\mu}_j$ for $j \in \{0, 1, 2\}$. Note that the underlying genetic model is identifiable by simply comparing these distributions. For example, a dominant model has overlapping distributions for $j = 1$ and $j = 2$ that are distinct from the distribution of $\hat{\mu}_0$.

Differences in the $\hat{\mu}_j$ indicate the SNP has an association with the phenotype. To determine whether any differences between the mean effect of the SNP levels are significant, we need to define an appropriate hypothesis test with the null and alternative hypotheses H_0 and H_1. The null and alternative hypotheses to test for any difference in the mean effects of different levels are as follows:

H_0: $\mu_0 = \mu_1 = \mu_2$ indicates that three mean effects are equal.
H_1: Not all the mean effects are equal.

As with classic GWAS, a primary assumption for the cell means model is that $\boldsymbol{Y_i}$ is normally distributed. Thus, this hypothesis test is done with F statistics, but inference in this setting is similarly impacted by violations of the normality assumption. Type-I error rates and power may be adversely impacted. Therefore, for the DCI–GWAS results we use the nonparametric Kruskal–Wallis test to assess the statistical significance of differences in the three distributions of $\hat{\mu}_j$. It is important to point out that we are proposing that our method is finding alternative sampling distributions for the mean effects of the different minor allele counts; thus, no testing on outcomes would be appropriate for this question. We did, though, confirm (results not shown) that our conclusions from testing the DCI–GWAS derived parameter distributions held even versus testing the DCI–GWAS derived updated (or even original) outcomes.

We summarize the three major elements of our methods as follows. First, we replace the classic-GWAS linear model with the cell means model to overcome the additive genetic. Second, we use DCI to estimate the distribution of the mean effect of each specific number of minor alleles. There is no normality assumption for DCI. Finally, we use nonparametric statistics to test for differences in these distributions to identify differential effects on the phenotype of the different numbers of minor alleles, hence, finding an association between the SNP and trait in an assumption-free model. The DCI algorithm is also summarized in a step-by-step visual after the real data COPDGene section; however, the simulation section describes the generation of our simulation data, and then we applied the DCI method as described in the visual.

4 | SIMULATION

In this section, we report on our comparison of the performance of the classic-GWAS additive model to our alternate DCI-estimated cell means model in a range of simulated scenarios. We compare our method to the GWAS additive model and not a cell-means linear model (such as analysis of variance) because we are proposing our approach as a replacement for what is currently most frequently used in practice. To be sure, though, we confirmed (not shown here) that our conclusions held even if we compared with classic GWAS calculated in the cell means framework. In addition, we compare Type-I error rates in a scenario where there is no difference in the mean effect on the trait for different minor allele counts for the SNP. We then provide power comparisons in five scenarios derived from our previous discussion of additivity and normality:

1. the genetic model is Additive and the normality of errors is *true*,
2. the genetic model is Additive and the normality of errors is *false*,

3. the genetic model is Non-Additive and the normality of errors is *false*,
4. dominant,
5. heterozygote advantage.

4.1 | Data simulation

The simulation framework is designed to assess the false-positive rate and the power for the two GWAS approaches in our real data setting using genotype data from the COPDGene project. The credibility of the data with respect to studying variant to phenotype maps has been repeatedly established by a range of domain experts, letting us be sure any findings were more likely true and not statistical artifact. More, those same domain experts believed there was more information in the data that novel methods might be able to exploit. We explain the study in more detail in the COPDGene section. However, for now the key detail is that we need to establish the credibility of our method on a roughly standardized continuous outcome for a biallelic SNP with minor allele frequencies of at least 5% (and often much larger). COPDGene also had approximately 10,000 participants in total. From the point-of-view of DCI, covariates are treated as constants simply adjusting the related observations; thus, we focus only on the SNP effects and residual errors here. In our COPDGene replication we adjust the two traits we are studying for necessary covariates and explain why this is a valid extension of our simulation conclusions. The model has the form for $i = 1, ..., 10,000$:

$$Y_i = \beta_0 + X_{i,SNP}\beta_{SNP} + \epsilon_i,$$
$$\epsilon_i \sim N(0, \sigma^2).$$

Our initial step in creating our synthetic data is to randomly generate a vector of length 10,000 of minor allele counts for an SNP or $\boldsymbol{X_{SNP}}$. This results in $\boldsymbol{X_{SNP}} = (X_1, ..., X_{10,000})^T$, where X_i is the count of minor alleles for participant i. Specifically, this is done in two stages. First, a random value is generated from a uniform distribution from 5% to 50% to use as the minor allele frequency (MAF). Next, this frequency was used to simulate an allele count (0, 1, or 2) for each of the 10,000 simulated participants. Technically, the allele count was generated from a binomial distribution for 2 outcomes with this MAF as the probability.

The next step is to simulate the outcome or trait for each of these 10,000 participants. These Y_j are simulated from the classic-GWAS model stated above in this section. The β_{SNP} is one of three values of β_{SNP_j} depending on the SNP level (j) for participant j. The β_{SNP_j} are chosen to match the additivity or nonadditivity of the desired genetic model. For example, in the additive case β_{SNP_j} is the same for $j \in \{0, 1, 2\}$, but in the dominant setting $\beta_{SNP_1} = \beta_{SNP_2}$ and these are both greater than β_{SNP_0}. Moreover, to have challenging simulations, we set the difference in β_{SNP_j}s small to about $\frac{1}{3}$ of the standard deviation in the real data. Our real COPDGene data outcomes were on residuals (that accounted for the covariates like race and ethnicity PCAs, gender, etc.), and simulation outcomes are treated as centered as the real data example. Therefore we used SDs of the real data to describe β_{SNP_j}s. Because our real data is standardized, we assume $\beta_0 = 0$ in our simulations. The specific values for our β_{SNP_j} in our five scenarios are given in the detailed descriptions of the scenarios.

Last, we simulate ϵ_j from a normal or slightly skewed distribution depending on whether our scenario assumes the normality of the errors is *true* or *false*. When we apply DCI to find our $\hat{\beta}_{SNP}$s distribution, we are implicitly assigning a variance to this distribution. Consequently, the variance of the residuals is an important element to consider in the context of the variance related to $\hat{\beta}_{SNP}$. Without an error term in our simulated outcomes, the trait would be completely determined by the SNP. If the SNP is truly associated with the outcome, though, the magnitude of the errors should be minimal (or the error would dominate the predicted Y causing any estimation approach to fail). To balance these concerns, we made the initial standard deviation of our residuals 0.35 or approximately 20% of our outcome's standard deviation. In this way, we give the errors a meaningful variance while still linking the majority of the changes in our outcomes to the SNP values and their possible effect sizes. Because we want to quantify power when there is in fact an SNP–phenotype association, our simulations have to allow this. We checked the robustness of this choice, and the values were stable for values smaller and larger than 0.35.

Each scenario was repeated with new random variables 200 times. As appropriate, we provide the Type-I error rate or the power and range of p values for the 200 iterations of each scenario. False and true positives are based on the GWAS threshold of 5×10^{-8}. All data generation and analysis were conducted using Python (version 3.8) and `numpy`, `pandas`, `sklearn.linear_model`, `statsmodels.api`, `scipy.stats`, and `matplotlib.pyplot` packages. We also used `ggplot2` from RStudio (version 2021.09.2) to create Miami plots.

4.2 | Simulation scenarios

The first scenario is to consider additive cases that deviate from normality. Figure 1 shows an example of our first simulation scenario.

In Figure 1, $Y0$ are the outcomes for those with 0 minor alleles. $Y1$ and $Y2$ are the outcomes for those with 1 and 2 minor alleles, respectively. There is an increasing and uniform shift in the distributions from $Y0$ to $Y1$ to $Y2$ indicating the additive assumption holds. However, the distributions of the outcomes are not normally distributed as we generated our mean effects from a very slightly skewed normal distribution. Moreover, observe that there is severe overlap in the distributions. This speaks to our motivation: can we update GWAS to find these shifts (whether additive or not) in the presence of uncertainty (whether normally distributed or not)?

To evaluate the classic-GWAS model, we combine all three different simulated outcomes $\mathbf{Y_0}$, $\mathbf{Y_1}$, and $\mathbf{Y_2}$ into a single continuous outcome vector $\mathbf{Y}$ and conduct the least squares regression under the additive model assumption. That is, we regress $\boldsymbol{Y}$ onto $\boldsymbol{X_{SNP}}$, and perform the hypothesis test for an association. As evident in Figure 2, the line that represents the additive association between the SNP and the observed outcome $\mathbf{Y}$ gives limited information about the association between the observed data and the SNP vector $\boldsymbol{X_S}$. Our simulations will show that the DCI adaptation can find this association in a comparable or better way than classic GWAS.

The DCI–GWAS is done in stages. We partition the data (simulated SNPs and outcomes) into three groups depending on the minor allele count. Within each partition, we then apply DCI as described in Section 3 to find a distribution for each μ_j. Our assumed initial distributions are $Normal(0, 0.35^2)$ for the errors (as described above) and $Normal(\bar{Y}_j, SE^2(Y_j))$ for μ_j. The second assumption implies that we are assuming that the sampling distribution for $\hat{\mu}_j$ found with least squares is correct. This will only change if it is inconsistent with the actual observed data. DCI is based on creating the range and relative frequency of outcomes based on possible values of μ_j. Thus, we sampled our initial distributions equal to the number of 0s for μ_0, the number of 1s for μ_1 and the number 2s μ_2 to have a sufficiently accurate approximation to $Normal(\bar{Y}_j, SE^2(Y_j))$. DCI uses the resulting predictions from this set of sampled parameters (specifically their consistency with the observed outcomes) to update the initial parameters and then use these to create an updated set of predictions.

To see an illustration of this, refer to Figure 3. The three panels show the observed, predicted based on the initial parameter distribution assumptions, and DCI-derived

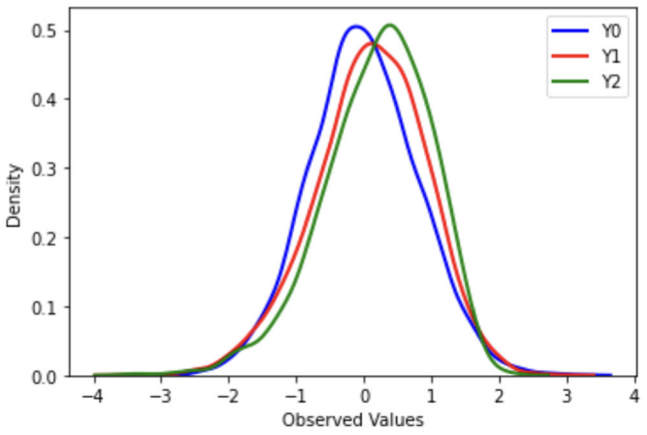

FIGURE 1 An example of the Additive Non-Normal scenario. $Y0$, $Y1$, and $Y2$ are the outcomes for those with 0, 1, and 2 minor alleles, respectively.

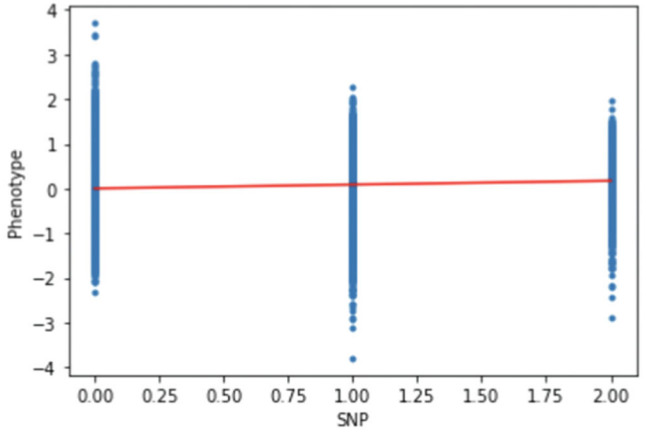

FIGURE 2 Regressing observed outcome $\mathbf{Y}$ onto $\boldsymbol{X_{SNP}}$ under the additive model assumption. SNP, single-nucleotide polymorphism.

updated outcomes for each minor allele count. For this figure, the red curve is the density for the true or observed outcomes. The blue curve is the density for the predicted outcomes or the push-forward of the initial. The dotted black curve is the density of the updated outcomes generated using DCI for each SNP level. The initial distribution in blue is different than the true red curve, indicating that the initial parameter distributions do not accurately predict the observed outcome. DCI reweights the parameters' initial distributions based on the degree of inconsistency between the observed outcomes and the predictions resulting from different values of the initial distribution. The dotted black densities, the predictions from the updated parameter distributions, are now near-perfect matches with the true.

When DCI updates the predicted outcomes $\hat{\boldsymbol{Y}}_{\boldsymbol{j}}$, it also updates all the parameters included in the model like μ_j and $\boldsymbol{\epsilon_j}$, keeping the parameter values that generate predictions consistent with the observed outcomes while

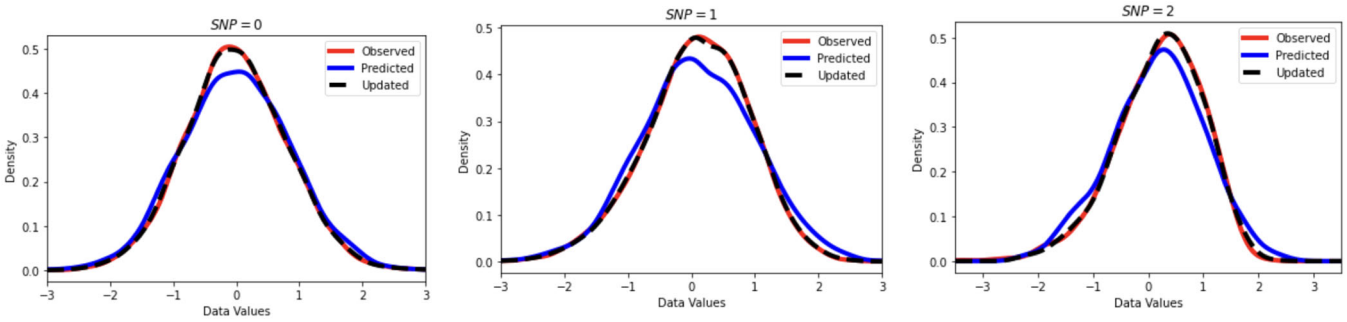

FIGURE 3 Observed, predicted based on the initial parameter distribution assumptions, and DCI-derived updated outcomes for each minor allele count. The red curve is the density of the observed outcomes. The blue curve is the density for the predicted outcomes. The dotted black curve is the density of the updated outcomes generated using DCI for each SNP level. DCI, data-consistent inversion; SNP, single-nucleotide polymorphism.

rejecting parameter values that generate predictions inconsistent with the observed outcome. If the updated predictions are consistent with the true outcomes, then the updated parameter distributions that caused these predictions are a viable density-weighted set of estimates of μ_j; that is, a data-consistent distribution for $\hat{\mu}_j$. Figure 4 depicts plots of initial and updated distributions for the mean effect of the SNPs for each minor allele count, where the blue curve is the initial distribution and the red curve is the DCI-derived updated distribution for each μ_j. Notice that the updated distributions reflect how the DCI distributions are minor (but important) data-adaptive changes from the initial distributions, which were based on the classic-GWAS assumptions were true. This illustrates how DCI can use the data to adapt to needed deviations from normality assumptions.

Our official test for an association was done by applying the Kruskal–Wallis test to these three distributions μ_0, μ_1, and μ_2. We conclude there is an association between the SNP and outcome is the resulting p value is less than the GWAS threshold of 5×10^{-8}.

We explored four more simulation scenarios. All used the standard deviation of 0.35 for the residuals. Note, below we provide the μ_j for each scenario, and 0.35 is high in comparison to these values. This forced substantial overlap in our scenarios. The overlap presents added challenges in finding associations, but it also better matches distributions in real data. The four scenarios are pictured in Figure 5. In all the scenarios, the deviations from normality are slight to ensure that our method works in difficult settings (especially ones where the deviation is not clear through observation so a researcher would not use a t test anyway).

Plot **(a)** depicts the additive and normal scenario with high SD. On the basis of the definition of the additive model (for **A**), there is a uniform, linear increase in risk for each copy of the **A** allele; the risk of having two copies of the **A** (**AA**) is twice the risk of having one copy of the **A** (**Aa**). In this case, we simulated the observed outcomes $\boldsymbol{Y_0}$, $\boldsymbol{Y_1}$, and $\boldsymbol{Y_2}$ from a normal distribution with a mean of zero, one, and two, corresponding to the *MAF* of 0, 1, and 2, respectively.

Plot **(b)** shows a case scenario where both normality and additivity assumptions are violated. To depict this case, we simulated 200 iterations with $\boldsymbol{Y_0}$s from a right-skewed normal distribution with mean 0, $\boldsymbol{Y_1}$s from a left-skewed normal with mean 0.1, and $\boldsymbol{Y_2}$s from a left-skewed normal with mean 0.15 to break normality and additivity assumptions. One subtle point is that in our simple sampling approach, stratified by allele count, non-normality in the outcomes directly indicates non-normality in the errors.

The dominant case scenario is shown in Figure 5 part **(c)**. A dominant model (for **A**) assumes that having one or more copies of the **A** allele increases risk compared with **a** (i.e., **aA** or **AA** genotypes have higher risk). For this scenario, we generate simulated observed outcomes $\boldsymbol{Y_0}$ from a right-skewed normal with mean 0, $\boldsymbol{Y_1}$ from a left-skewed normal with mean 0.1, and $\boldsymbol{Y_2}$ from a left-skewed normal with mean 0.11, corresponding to the *MAF* of 0, 1, and 2 to keep the model dominant.

For that last case scenario, we consider the heterozygote advantage case scenario shown in Figure 5 part (d). In this special and rare case, having one minor allele count ($MAF = 1$) increases the risk of a special phenotype while having two minor allele counts ($MAF = 2$) decreases the risk of a special phenotype or has a protective effect. For this scenario the observed outcomes $\boldsymbol{Y_0}$, $\boldsymbol{Y_1}$, and $\boldsymbol{Y_2}$ were simulated from a slightly right-skewed normal distribution with mean zero, a left-skewed normal distribution with mean -0.1, and a right-skewed normal distribution with mean 0.1, corresponding to the *MAF* of 0, 1, and 2, respectively.

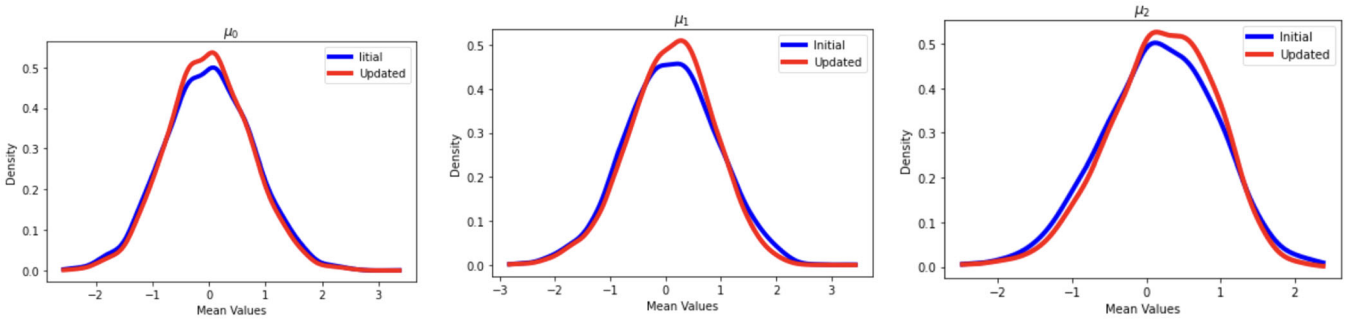

FIGURE 4 Initial and updated distributions for the mean effect of the SNPs, μ_js, for each minor allele count. The blue curve is the initial distribution and the red curve is the DCI-derived updated distribution for each μ_j. DCI, data-consistent inversion; SNP, single-nucleotide polymorphism.

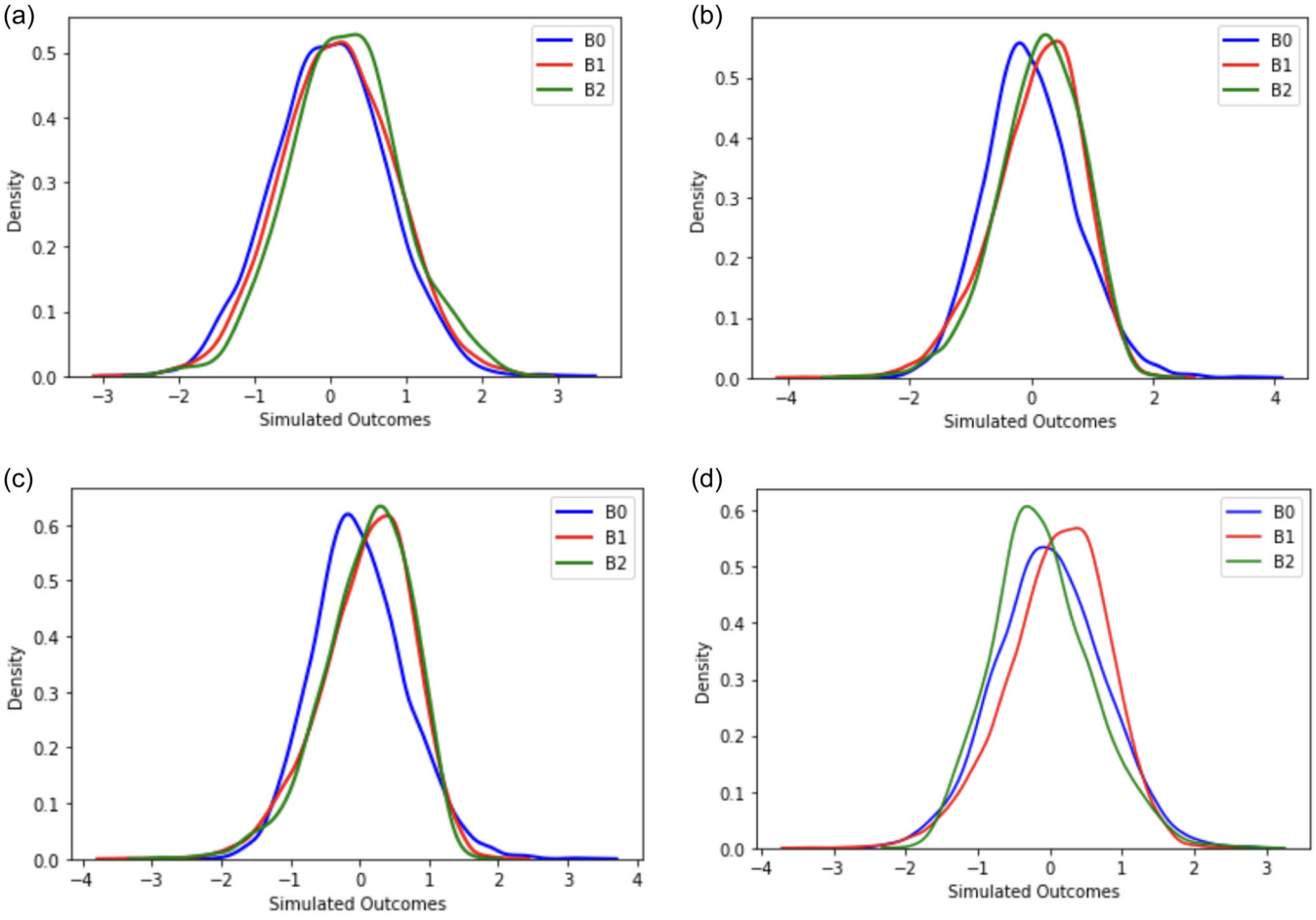

FIGURE 5 Examples of each simulation scenario. In each plot, $Y0$, $Y1$, and $Y2$ are the outcomes for those with 0, 1, and 2 minor alleles, respectively. Plot (a) meets both additive and normal assumptions. Plot (b) deviates from both normality and additivity. Plot (c) depicts a dominant case scenario. Plot (d) shows a heterozygote advantage case scenario.

Finally, we generated outcomes from the simulation scenarios except where the null assumption of no variant effect is true (no difference by allele count). We used these to calculate the rate of rejecting the null hypothesis to assess the Type-I error rate.

4.3 | Simulation results

With respect to false positives, classic (regression) GWAS and our DCI–GWAS performed identically. Moreover, the false rejection rate (Type-I error rate) was less than 0.05.

In Table 1, we report details about the power of the two approaches in our five scenarios. The scenario is in the left-most column. The results for the classic GWAS are in the next two columns, and the results for our DCI-based approach are in the last two columns. The power, provided as the percent of the 200 iterations of the scenario in which we conclude there is an association between the SNP and outcome, is given in the columns labeled "Association." The ranges of p values from the 200 tests (t test for GWAS and Kruskal–Wallis for DCI) are shared in the "p value" columns.

The new DCI–GWAS had at least 90% in all scenarios, whereas the classic GWAS had noticeably less than this level of power for the "Dominant" and "Heterozygote Advantage" scenarios. As the recessive case is symmetric to the dominant case with respect to how DCI estimates the mean effects, we believe this conclusion is true for the recessive case as well. Our approach is able to adapt to all the underlying genetic models, but the classic GWAS is, as the literature suggested, not able to similarly adapt.

Next, there is a pattern of how the performance of DCI–GWAS compares to the classic GWAS. In the setting that meets the linear model or least squares regression assumptions, where additivity and normality are true, both methods had extremely good power but the classic GWAS had 1.5% more power than DCI–GWAS. This is encouraging because it shows that DCI–GWAS performs comparably to the classic GWAS even in the ideal scenario for classic GWAS. More, the DCI approach is using a nonparametric statistic in a situation where the parametric is known to be valid. Thus, it is just as encouraging that the power only decreased by such a small amount (1.5%).

When the scenarios have even one violation of the assumptions, then DCI–GWAS has superior power. If there is non-normality of the errors, DCI–GWAS was slightly better (1%) even if the underlying model was additive. If the additivity was also invalid, the advantages of DCI–GWAS were stronger. In the third scenario, the model is neither additive nor dominant, and our method outperformed classic GWAS by 5%. In the dominant scenario, the improvement grew to nearly 20%. Finally, in our unique Heterozygote Advantage scenario, we confirmed that classic GWAS is unable to capture this type of association, although our adapted version was successful in more than 90% of our iterations; that is, we attained at least 90% power.

Although the $10,000$ sample size matched the COPDGene data, we repeated the simulation with $n = 1,000,000$ to show the reproducibility of the method to more current GWAS sizes. Table 2 summarizes the results for $1\,M$ cases. The results show approximately the same power as the results for $10,000$ sample size and even more power in some of the scenarios. In all scenarios, we had at least 90% power like the $10,000$ sample size. The classic GWAS had noticeably less power for the "Dominant" and "Heterozygote Advantage" scenarios for $1\,M$ cases as well. Furthermore, the same pattern in the performance of DCI–GWAS compared with the classic GWAS is repeated here in $n = 1,000,000$. This shows DCI–GWAS outperforms

TABLE 1 Simulation results: The power of the classic-GWAS and DCI-based approach in five scenarios for 10K simulated cases.

Number of cases = 10,000	Number of iterations = 200			
	GWAS		DCI	
Simulation cases	Association (%)	p Value	Association (%)	p Value
Additive normal	97.5	>1.68E − 36	96	>1.18E − 37
		<8.49E − 07		<9.06E − 05
Additive non-normal	91.5	>2.28E − 36	92.5	>6.22E − 42
		<5.33E − 05		<3.19E − 03
Nonadditive non-normal		>5.38E − 33	94	>1.47E − 37
		<5.99E − 05		<1.86E − 04
Dominant	74.5	>1.08E − 27	93	>3.06E − 36
		<9.29E − 04		<1.42E − 04
Heterozygote advantage	0	>7.38E − 8	90.5	>2.05E − 30
		<9.81E − 1		<3.09E − 04

Note: "Association" columns show the power provided as the percent of the 200 iterations of the scenario in which an association between the SNP and outcome was concluded. "p Value" columns share the ranges of p values from the 200 tests (t test for GWAS and Kruskal–Wallis for DCI).

Abbreviations: DCI, data-consistent inversion; GWAS, genome-wide association study; SNP, single-nucleotide polymorphism.

the classic GWAS in more scenarios in the larger sample sizes typically seen in GWAS today.

The allele frequencies used in the simulations were chosen to allow the results to generalize to a wide range of potential real scenarios. That is, we chose a large enough number of simulation iterations (200) to test the method robustly throughout the 0.05 to 0.50 interval of Minor Allele Frequencies, an interval which encompasses many true SNP allele frequencies. In addition to this, we reran all simulations with an allele frequency of exactly 0.05 and then exactly 0.50. The results were the same as those provided in the Simulation Results Table 1, confirming that are conclusion held for the complete interval of allele frequencies in $[0.05, 0.50]$. Although it was outside the goal of our developing a method for common variants, we still tested the method with an $MAF = 0.01$; unfortunately, the method performed insufficiently for us to recommend using it for rare alleles.

To complete our results, Table 3 shows the rates of false positives for the different distribution scenarios across the classic and DCI versions of GWAS. For consistency, the results are based on the same five simulation scenarios used above except that the null hypothesis is now assumed to be true. That is, we used the same simulation setup except with zero difference in the effect of the outcome whether the allele count was 0, 1, or 2. Encouragingly, the DCI method controlled the Type-I error rate at the desired 0.05 threshold for all scenarios, even exhibiting evidence of being slightly conservative in the dominant and heterozygote advantage settings. There was also no inflation in the rate for the DCI-based GWAS compared with the traditional GWAS approach. In fact, the rate for the DCI approach was slightly lower in all but the additive normal setting (the optimal one for classic GWAS). As there was no inflated false-positive rate for the DCI–GWAS compared with classic GWAS, the differences in the power results are even more likely real and not driven by an understated true Type-I error rate.

When removing assumptions from a model, there is always concern about a trade-off in the model's performance regarding statistical inference. Our simulation experiment demonstrated that removing the normality and additivity assumptions may not reduce our model's power. In the perfect scenario where the true genetic structure is additive and the errors are exactly normal, the classic GWAS may still be the best choice. However, even in this setting the power trade-off for our method was approaching negligible. In our other scenarios, though, our DCI adaptation of GWAS outperformed the classic GWAS, sometimes by a significant amount. Considering that our method is agnostic to the underlying genetic structure and that real data is rarely perfectly additive or normal, the simulations suggest that our GWAS tool has potential as an improvement over the classic version.

4.4 | Computation time

Across all the simulations, the changes in computational time for increased sample size were consistent. The time complexity for the DCI–GWAS method increases quadratically as a function of sample size.

TABLE 2 Simulation results for 1*M* cases: The power of the classic-GWAS and DCI-based approach in five scenarios for 1*M* cases.

Number of cases = 1,000,000	Number of iterations = 200			
	GWAS		DCI	
Simulation cases	Association (%)	*p* Value	Association (%)	*p* Value
Additive normal	97.5	>2. 45E − 38	96	>1.39E − 38
		<1. 09E − 07		<1.03E − 05
Additive non-normal	91.5	>9.34E − 37	93	>1.13E − 43
		<2.53E − 05		<1.09E − 03
Nonadditive non-normal	89	>1.02E − 33	95	>1.76E − 38
		<8.21E − 04		<5.23E − 05
Dominant	73.5	>7.02E − 26	93.5	>1.02E − 37
		<1.39E − 03		<5.23E − 03
Heterozygote advantage	0	>1.06E − 7	91	>1.31E − 31
		<4.36E − 1		<1.02E − 04

Note: "Association" columns show the power provided as the percent of the 200 iterations of the scenario in which an association between the SNP and outcome was concluded. "*p* Value" columns share the ranges of *p* values from the 200 tests (*t* test for GWAS and Kruskal–Wallis for DCI).

Abbreviations: DCI, data-consistent inversion; GWAS, genome-wide association study; SNP, single-nucleotide polymorphism.

TABLE 3 False-positive rate: The false-positive discovery rates for the different distribution scenarios across the classic and DCI versions of GWAS.

Simulation cases	GWAS	DCI
Additive normal	0.045	0.05
Additive non-normal	0.05	0.047
Nonadditive non-normal	0.051	0.046
Dominant	0.054	0.04
Heterozygote advantage	0.067	0.038

Note: The results are based on the same five simulation scenarios.

Abbreviations: DCI, data-consistent inversion; GWAS, genome-wide association study.

5 | APPLICATION TO COPDGENE GWAS

Chronic obstructive pulmonary disease (COPD) was the second most common cause of death (in terms of age-standardized death rate) across the world in 1990. An increase of 15.6% in the prevalence of COPD was reported from 2007 to 2017 (Safiri et al., 2022). Now, it is the third leading cause of death worldwide, causing 3.23 million deaths in 2019 reported by WHO in May 2022 (WHO, n.d).

COPDGene is a multisite, longitudinal study with the goal of identifying genetic and clinical determinants of chronic pulmonary obstructive disease (COPD, n.d). The study has been described in more detail previously (Regan et al., 2011), including the study design and characteristics. The processing of the data, the collection methods, and informed consent descriptions are available in the cited study design paper. One of the most significant findings is that COPD is not one homogeneous disease, the long-held consensus by previous researchers and clinicians. The COPDGene data was used to identify COPD subtypes particularly emphysema predominant disease (EPD) and airway-predominant disease (APD) axes (continuous variables representing the strength of these subtypes) (Kinney et al., 2018). To validate this finding, these COPD subtypes identified individuals at risk for mortality (Young et al., 2019). The subtypes were identified from linear combinations of observable (measurable) variables related to a patient's pulmonary function, inspiratory and expiratory computerized tomography, and airway measurements (Kinney et al., 2018). Genetic information was not used in identifying patients with the EPD and APD subtypes of COPD.

Consequently, the natural next research goal was to identify genetic variants associated with EPD and APD. Young et al. (2016) performed the related GWAS, finding SNPs associated on the *AGER* gene for EPD and on *CHRNA*5/3/84 for APD within a Non-Hispanic White (NHW) population (Young et al., 2016, 2019). These GWAS, using the classic additive linear model approach, also uncovered a number of marginal candidates for the subtypes.

Our motivation is to apply our new DCI–GWAS to this particular data set for multiple reasons. First, we want to generally replicate the findings of the original GWAS. Second, we want to see if we learn more about the marginal candidate SNPs from our tool; for example, our method may increase their significance if the most likely genetic model was recessive or dominant. Finally, we want to apply our tool in a setting (COPD) known to present challenges when modeling the SNP-phenotype maps; thus, we might learn beyond what our simulations revealed about when our method has advantages and when it does not. We will present the results using DCI–GWAS as well as the classic GWAS for comparison purposes.

COPDGene participants self-identified as either African American (AA) or NHW. As we described before, we perform our DCI–GWAS separately for these two populations to minimize potential heterogeneity in our residual variance. Note, DCI–GWAS could analyze the combined sample by using it to estimate the distributions of the variances in these two populations. Stratification is used to improve the classic-GWAS model, and we will use DCI–GWAS for the two groups individually to let us compare the two methods' performances. This solution will not address population stratification concerns, though. Thus, we adjusted for the first five principal components in our models. Finally, we included known confounders for COPD: age, gender, and smoking status (current or former). For computational ease and to align with the structure of our simulations, we conducted DCI–GWAS on the residuals from regressing EPD or APD onto this set of variables. That is, we adjusted the subtype outcomes for population stratification and covariates before performing the association testing through DCI–GWAS for the SNPs.

In total, our final samples included 2476 AA patients and 5526 NHW patients. The average age was 60, 46.3% were female, and 52.1% were current smokers. Our GWAS included 540,687 SNPs for the AA-APD GWAS, 540,687 SNPs for the AA-EPD GWAS, 509,854 SNPs NHW-APD GWAS, and 509,854 SNPs for the NHW-EPD GWAS.

5.1 | EPD and APD GWAS results

We applied our DCI-derived model to COPDGene data and created Miami plots to compare DCI-derived and classic-GWAS results for EPD and APD in AA, and NHW populations. The Miami plot is a mirrored or paired set of Manhattan plots, the top for DCI and bottom for classic

GWAS. In each the $-log_{10}(p$ value) for each GWAS SNP is on the vertical axis, and the SNPs themselves are organized by location (and chromosome) on the horizontal axis. The GWAS thresholds for a significant SNP-subtype association are the horizontal red lines (a suggestive line for 10^{-5} is also plotted).

Figure 6, the first Miami set of plots, is for the AA COPDGene patients. Only one SNP reaches GWAS significance, and only for EPD. There is no classic-GWAS comparison for our one significant SNP to establish a similar ability to detect correct associations. However, these GWAS provide a degree of

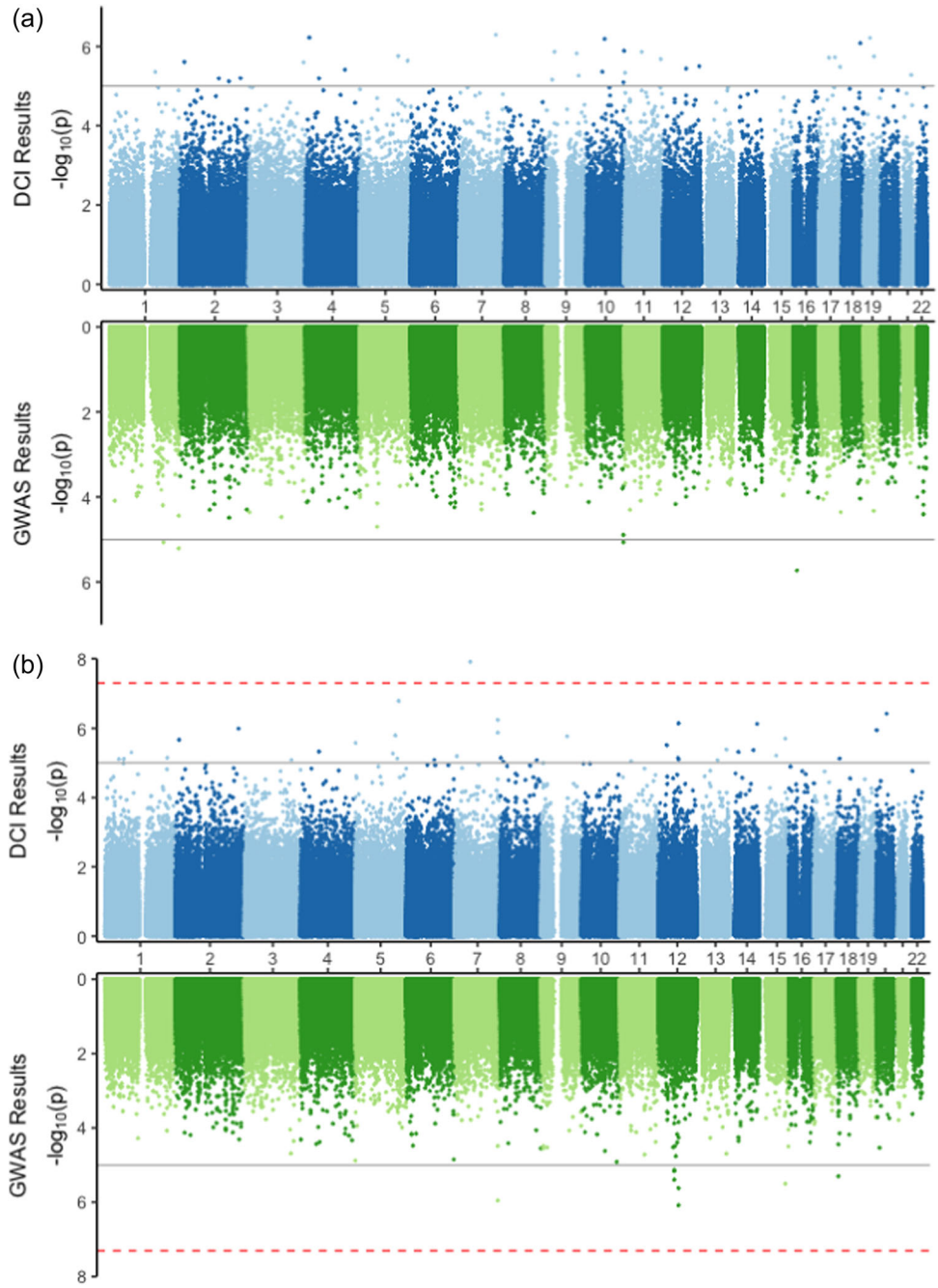

FIGURE 6 Miami plots for the African American COPDGene populations. Plot (a) shows the results of the APD and plot (b) depicts the results of the APD. The top panel shows the DCI-derived results, although the bottom panel shows GWAS results. In the Miami plot $-log_{10}(p)$ is plotted on the y-axis and chromosomal location is plotted on the x-axis. The genome-wide significance threshold ($p < 5 \times 10^{-8}$) is indicated by a red dashed line, although suggestive significance ($p < 5 \times 10^{-5}$) is indicated by a light gray line. (a) Miami plot for APD and (b) Miami plot for EPD. APD, airway-predominant disease; COPD, chronic obstructive pulmonary disease; DCI, data-consistent inversion; GWAS, genome-wide association study; EPD, emphysema predominant disease.

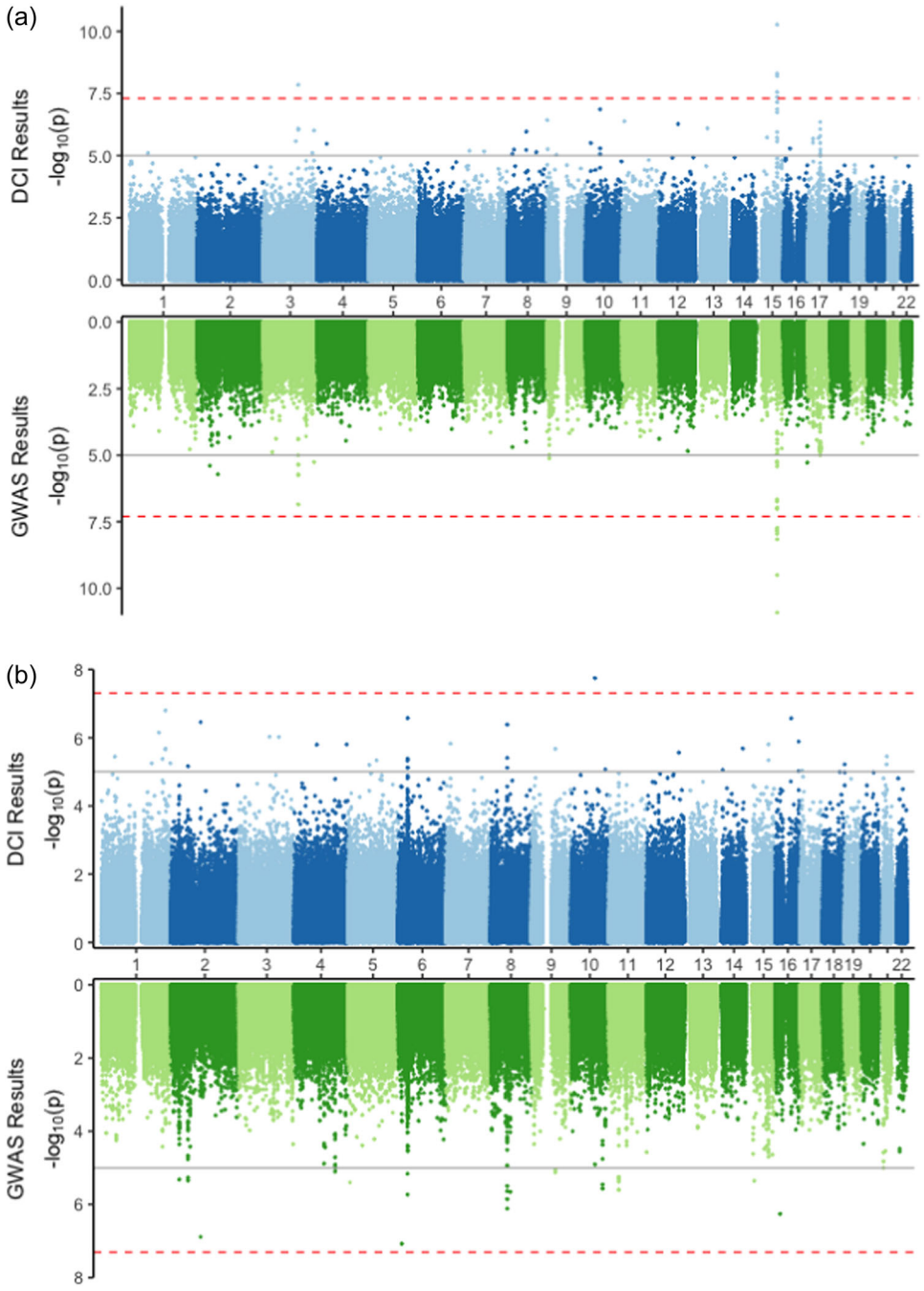

FIGURE 7 Miami plots for the Non-Hispanic White COPDGene populations. Plot (a) shows the results of the APD and plot (b) depicts the results of the APD. The top panel shows the DCI-derived results, although the bottom panel shows GWAS results. In the Miami plot $-log_{10}(p)$ is plotted on the y-axis and chromosomal location is plotted on the x-axis. The genome-wide significance threshold ($p < 5 \times 10^{-8}$) is indicated by a red dashed line, although suggestive significance ($p < 5 \times 10^{-5}$) is indicated by a light gray line. (a) Miami plot for APD and (b) Miami plot for EPD. APD, airway-predominant disease; COPD, chronic obstructive pulmonary disease; DCI, data-consistent inversion; GWAS, genome-wide association study; EPD, emphysema predominant disease.

confirmation that DCI–GWAS does not inflate the Type-I error rate.

Figure 7, the second set of Miami plots, provides the results of the APD and EPD in the NHW COPDGene populations. The EPD can be similarly interpreted as those of the AA results. The airway-predominant results allow several key insights into our method's potential.

Related to our primary hypotheses, first we replicated the majority of the classic-GWAS findings. These SNPs were all on chromosome 15, on the CHRNA3, CHRNA5, and HYKK genes. These three genes have been linked to

COPD and smoking phenotypes in multiple independent studies (Kaur-Knudsen et al., 2012; Nedeljkovic et al., 2018; Pérez-Morales et al., 2018). To be specific, DCI–GWAS reached the GWAS threshold for four of the eight SNPs, but the p value for two more was 7×10^{-8}, and 1×10^{-7} and 2×10^{-6} for the other two. Thus, we replicated approximately three-quarters of the classic-GWAS findings overall, and, importantly, found signals on the key genes.

There were three other noteworthy findings. First, our method identified an SNP on the IREB2 gene with a known association with Post bronchodilator Forced Expiratory Volume in 1 s (FEV_1), a trait related to airway disease (Lutz et al., 2015). Classic GWAS did not identify this SNP as GWAS significant. Second, our method identified an SNP on the EEFSEC gene in chromosome three for which classic GWAS was only at the 2×10^{-6} level. The EEFSEC gene has been linked to COPD in more recent literature (Benway et al., 2021; Hobbs et al., 2017). Further, the SNP was identified by a GWAS of GOLD Stage (a commonly used categorical classification of COPD severity) (Gold, n.d) in our same COPDGene cohort. Figure 8 shows the distributions of the airway disease trait by minor allele count. Notice the overlap of the trait values for those with 0 and 1 minor alleles, but a lower (more severe APD) distribution for those with two copies of the minor allele. This recessive model scenario as exactly our background literature and simulations, could be successfully modeled with our DCI method but difficult for classic GWAS. Remember that the discovery of the subtypes (such as EPD and APD) revealed that COPD was likely a collection of diseases and not a single disease. GOLD stage treats COPD like a single disease; thus, DCI–GWAS shows that our SNP is really affecting the APD subtype and not EPD. Therefore, although a patient may be exhibiting severe enough symptoms to be diagnosed by the GOLD stage, our results indicate therapies should target APD.

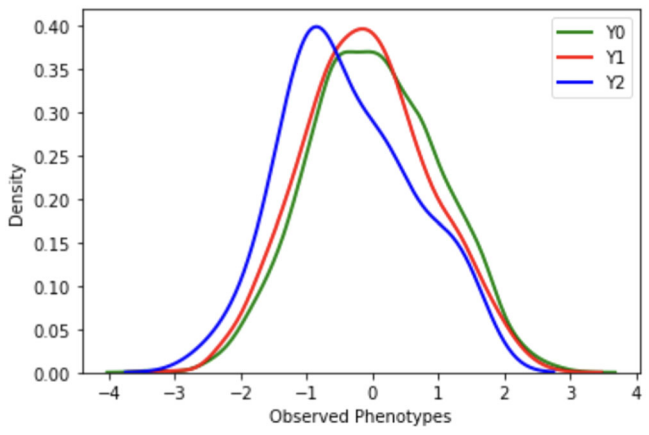

FIGURE 8 Distributions of the airway disease trait by minor allele count for *rs*2811544 SNP on the EEFSEC gene SNP on the EEFSEC gene in chromosome three. SNP, single-nucleotide polymorphism.

Overall, DCI–GWAS in the COPDGene populations confirmed the simulation findings. There was no evidence of inflating the rate of false positive. DCI–GWAS identified the majority of the SNPs found with classic GWAS and found SNPs in all the key genes identified by the classic GWAS. Perhaps most importantly, DCI–GWAS found two SNPs related to airway disease that did not reach the GWAS threshold with the classic approach. Both SNPs have been directly linked to COPD or COPD traits related to airway disease and both are on genes with known biologic function that would impact airway disease. Before moving to the Discussion, the following summarizes the steps used in this section.

6 | DISCUSSION

GWAS has led to enormous numbers of important discoveries that have enabled key scientific findings about genetic effects on traits. Despite its success, we hypothesized that its dependence on the correctness of the additive linear model it uses was limiting the potential for new GWAS discoveries. Further, although the linear model can be robust to some extent to violations of normality in the errors, there are times when the deviation is so subtle to be hard to see in a Q–Q plot or not fully remedied by a transformation. We hypothesized, that it is possible that in some of these situations, the slight deviation could cause classic approaches to miss true variant to disease associations. A method that was not dependent on the normality assumption would thereby have the potential to identify these associations and open up many new candidates for further study. In the current work, we replaced least squares estimation with a novel DCI (from the uncertainty quantification and computational mathematics domains) approach to remove this dependency. The result is a more generalized, genetic model agnostic, and normality assumptionless version of GWAS. Although we have found a way to remove many of the assumptions from GWAS, potentially allowing it to make even more discoveries, our method still has some limitations. Our method depends on a non-parametric statistic; thus, there are even more potential power gains possible with our method. We also have not optimized our results by determining the best balance of parameter and residual variance. We focused the current work on stable results, but future work could improve power by optimizing our approach. The method currently focuses only on genotype data as this lets us apply it to the promising COPDGene Project data. Future adaptations to imputed dosage data are an important next step. Similarly, as we are working in a linear model framework, future work

accounting for nonindependence in the residuals, such as that from related individuals, would be beneficial. One exciting feature of DCI is the ability to provide distribution estimates for any parameters in a model, as long as a credible model framework is established and reasonable initial assumptions about these distributions are possible. The appropriateness of the linear model with dosage has been established already in the literature; similarly, parameterizations of related individuals have been established. Thus, we are hopeful that our method will naturally extend in these settings.

Through simulation and replication of a COPDGene GWAS, we established the credibility of our method. It performs nearly identically to the classic GWAS even when the linear model assumptions are valid, but it outperforms when they are invalid. Of note, the new method was able to correctly identify the underlying genetic model without user input. In the COPDGene GWAS, the DCI-based GWAS found an SNP with a plausible effect on the COPD trait we were studying, and classic GWAS had not identified this SNP. Therefore, we are hopeful that our new GWAS approach can benefit those searching for new insights into how variants impact phenotypes, especially for COPD as it was shown to be successful using the COPDGene population. Even more, with our freely available code, applying the methodology to previous GWAS has exciting potential to use already existing data to identify or replicate even more genetic variants associated with diseases.

DCI procedure: Applying DCI method on COPDGean data.

Inputs:

- A vector of SNPs, $\boldsymbol{X_{SNP}}$.
- A vector of phenotypes or traits, **Y**.

Preprocessing computations:

1. Determine covariates of interest (Age, gender, smoking status, and first five PC).
2. Create a linear model with **Y** as the outcome and all covariates as variables.
3. Save the residuals as the vector of the observed outcome.
4. Separate $\boldsymbol{X_{SNP}}$ into three distinct vectors based on *MAF* count 0, 1, and 2 with their corresponding vectors of observed outcome.

Apply DCI method:

1. Make an initial normal distribution for parameter β_S with mean and standard deviation of vectors of observed outcome.

(Continues)

DCI procedure: Applying DCI method on COPDGean data.

2. Make an initial normal distribution with mean *zero* and a standard deviation of 0.35 as the initial error for each MAF count.
3. Add initial mean and initial error to create the vector of the predicted outcome.
4. Use GKDE to construct nonparametric distribution estimates of densities for both initial and observed outcomes.
5. Construct the ratio between the observed and predicted densities.
6. Use this ratio to reweight and update the initial distributions of the parameter and the predicted outcome.
7. Apply Kruskal–Wallis test on the updated parameters of each group.

Output: The p value of the difference between the mean of three *MAF* groups 0, 1, and 2.

ACKNOWLEDGMENTS

We thank Prof Jan Mandel, the director of the Center for Computational Mathematics, University of Colorado Denver, for providing the computing resources. We thank Dr. Michael Cho, Brigham and Women's Hospital, Harvard Medical School, for comments on the practical applicability of the method. We thank the University of Colorado Laboratory for Analytical and Computational Epidemiology (LACE) group for providing a multidisciplinary discussion group and support. This work was supported by NHLBI U01 HL089897 and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. Prof Butler's work is supported by the National Science Foundation under Grant No. DMS-2208460 as well as by the NSF IR/D program, while working at National Science Foundation. However, any opinion, finding, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from COPDGene. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of COPDGene.

ORCID

Negar Janani http://orcid.org/0009-0000-7308-926X
Kendra A. Young http://orcid.org/0000-0002-8664-7635

REFERENCES

Basu, S., Zhang, Y., Ray, D., Miller, M. B., Iacono, W. G., & McGue, M. (2013). A rapid gene-based genome-wide association test with multivariate traits. *Human Heredity*, *76*(2), 53–63.

Behrens, G., Winkler, T. W., Gorski, M., Leitzmann, M. F., & Heid, I. M. (2011). To stratify or not to stratify: Power considerations for population-based genome-wide association studies of quantitative traits. *Genetic Epidemiology*, *35*(8), 867–879.

Benway, C. J., Liu, J., Guo, F., Du, F., Randell, S. H., Cho, M. H., Silverman, E. K., Zhou, X., & Consortium, I. C. G. (2021). Chromatin landscapes of human lung cells predict potentially functional chronic obstructive pulmonary disease genome-wide association study variants. *American Journal of Respiratory Cell and Molecular Biology*, *65*(1), 92–102.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822.

Butler, T., Wildey, T., & Yen, T. Y. (2020). Data-consistent inversion for stochastic input-to-output maps. *Inverse Problems*, *36*(8), 085015.

Connor, G., & O'Neill, M. (2017). Finite-sample genome-wide regression p-values (GWRPV) with a non-normally distributed phenotype. *bioRxiv*, 204727.

COPD. (n.d.). COPD Genetic Epidemiology. http://www.copdgene.org

Fayers, P. (2011). Alphas, betas and skewy distributions: Two ways of getting the wrong answer. *Advances in Health Sciences Education*, *16*(3), 291–296.

Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *The American Journal of Human Genetics*, *71*(2), 217–222.

Fernando, R. L., & Garrick, D. (2013). Bayesian methods applied to GWAS. *Genome-wide association studies and genomic prediction* (pp. 237–274). Springer.

Goh, L., & Yap, V. B. (2009). Effects of normalization on quantitative traits in association test. *BMC Bioinformatics*, *10*(1), 1–8.

Gold. (n.d.). Global Initiative for Chronic Obstructive Lung Disease. https://goldcopd.org

Hobbs, B. D., De Jong, K., Lamontagne, M., Bossé, Y., Shrine, N., Artigas, M. S., Wain, L. V., Hall, I. P., Jackson, V. E., Wyss, A. B., London, S. J., North, K. E., Franceschini, N., Strachan, D. P., Beaty, T. H., Hokanson, J. E. Crapo, J. D., Castaldi, P. J., Chase, R. P., ... International COPD Genetics Consortium. (2017). Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nature Genetics*, *49*(3), 426–432.

Kaur-Knudsen, D., Nordestgaard, B. G., & Bojesen, S. E. (2012). CHRNA3 genotype, nicotine dependence, lung function and disease in the general population. *European Respiratory Journal*, *40*(6), 1538–1544.

Kinney, G. L., Santorico, S. A., Young, K. A., Cho, M. H., Castaldi, P. J., San JoséEstépar, R., Ross, J. C., Dy, J. G., Make, B. J., Regan, E. A., Lynch, D. A., Everett, D. C., Lutz, S. M., Silverman, E. K., Washko, G. R., Crapo, J. D., Hokanson, J. E., & COPDGene Investigators. (2018). Identification of chronic obstructive pulmonary disease axes that predict all-cause mortality: The COPDGene study. *American Journal of Epidemiology*, *187*(10), 2109–2116.

Konietschke, F., Libiger, O., & Hothorn, L. A. (2012). Nonparametric evaluation of quantitative traits in population-based association studies when the genetic model is unknown. *PLoS ONE*, *7*(2), e31242.

Lourenço, V. M., Pires, A. M., & Kirst, M. (2011). Robust linear regression methods in association studies. *Bioinformatics*, *27*(6), 815–821.

Lutz, S. M., Cho, M. H., Young, K., Hersh, C. P., Castaldi, P. J., McDonald, M.-L., Regan, E., Mattheisen, M., DeMeo, D. L., Parker, M., Foreman, M., Make, B. J., Jensen, R. L., Casaburi, R., Lomas, D. A., Bhatt, S. P., Bakke, P., Gulsvik, A., Crapo, J. D., ... ECLIPSE Investigators, and COPDGene Investigators. (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genetics*, *16*(1), 1–11.

Monir, M., Zhu, J. (2017). Comparing GWAS results of complex traits using full genetic model and additive models for revealing genetic architecture. *Scientific Reports*, *7*(1), 1–12.

Murabito, J. M., Rosenberg, C. L., Finger, D., Kreger, B. E., Levy, D., Splansky, G. L., Antman, K., & Hwang, S.-J. (2007). A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham heart study. *BMC Medical Genetics*, *8*(1), 1–10.

Nedeljkovic, I., Carnero-Montoro, E., Lahousse, L., van der Plaat, D. A., de Jong, K., Vonk, J. M., van Diemen, C. C., Faiz, A., Van Den Berge, M., Obeidat, M., Bossé, Y., Nickle, D. C., BIOS Consortium, Uitterlinden, A. G., van Meurs, J. J. B., Stricker, B. C. H., Brusselle, G. G., Postma, D. S., Boezen, H. M., ... Amin, N. (2018). Understanding the role of the chromosome 15q25. 1 in COPD through epigenetics and transcriptomics. *European Journal of Human Genetics*, *26*(5), 709–722.

Pérez-Morales, R., González-Zamora, A., González-Delgado, M. F., Calleros Rincon, E. Y., Olivas Calderon, E. H., Martínez-Ramírez, O. C., & Rubio, J. (2018). Chrna3 rs1051730 and chrna5 rs16969968 polymorphisms are associated with heavy smoking, lung cancer, and chronic obstructive pulmonary disease in a Mexican population. *Annals of Human Genetics*, *82*(6), 415–424.

Pires, A. M., & Rodrigues, I. M. (2007). Multiple linear regression with some correlated errors: Classical and robust methods. *Statistics in Medicine*, *26*(15), 2901–2918.

Ray, D., & Chatterjee, N. (2020). Effect of non-normality and low count variants on cross-phenotype association tests in gwas. *European Journal of Human Genetics*, *28*(3), 300–312.

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., Curran-Everett, D., Silverman, E. K., & Crapo, J. D. (2011). Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, *7*(1), 32–43.

Safiri, S., Carson-Chahhoud, K., Noori, M., Nejadghaderi, S. A., Sullman, M. J., Heris, J. A., Ansarin, K., Mansournia, M. A., Collins, G. S., Kolahi, A.-A., & Kaufman, J. S. (2022). Burden of chronic obstructive pulmonary disease and its attributable risk factors in 204 countries and territories, 1990–2019: Results from the global burden of disease study 2019. *BMJ*, *378*.

Schifano, E. D., Li, L., Christiani, D. C., & Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics*, *92*(5), 744–759.

Sitlani, C. M., Rice, K. M., Lumley, T., McKnight, B., Cupples, L. A., Avery, C. L., Noordam, R., Stricker, B. H., Whitsel, E. A., & Psaty, B. M. (2015). Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. *Statistics in Medicine*, *34*(1), 118–130.

Tran, A., & Wildey, T. (2021). Solving stochastic inverse problems for property–structure linkages using data-consistent inversion and machine learning. *JOM*, *73*(1), 72–89.

Verhulst, B., & Neale, M. C. (2021). Best practices for binary and ordinal data analyses. *Behavior Genetics*, *51*(3), 204–214.

WHO. (n.d.). World Health Organization. https://www.WHO.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

Young, K. A., Kinney, G. L., Cho, M. H., Castaldi, P. J., Beaty, T. H., Crapo, J. D., Silverman, E. K., Hokanson, J., & Lutz, S. M. (2016). Genome-wide association analysis of pulmonary disease factors reveals different genetic variation for emphysema and airway disease. In *B105 COPD: Epidemiology and genetics* (p. A4430). American Thoracic Society.

Young, K. A., Regan, E. A., Han, M. K., Lutz, S. M., Ragland, M., Castaldi, P. J., Washko, G. R., Cho, M. H., Strand, M., Curran-Everett, D., Beaty, T. H., Bowler, R. P., Wan, E. S., Lynch, D. A., Make, B. J., Silverman, E. K., Crapo, J. D., Hokanson, J. E., Kinney, G. L., & COPDGene® Investigators. (2019). Subtypes of COPD have unique distributions and differential risk of mortality. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation*, *6*(5), 400.

Zachary, C., Young, K., Kinney, G., Strand, M., Ragland, M., Janani, N., Li, Y., Richmond, N., Hokanson, J., Butler, T., & Austin, E. (2022). Using data consistent inversion to describe and quantify the impact of sources of uncertainty on predictions of lung function in the COPDGene cohort. *Population Health Metrics*.

How to cite this article: Janani, N., Young, K. A., Kinney, G., Strand, M., Hokanson, J. E., Liu, Y., Butler, T., & Austin, E. (2024). A novel application of data-consistent inversion to overcome spurious inference in genome-wide association studies. *Genetic Epidemiology*, 1–19. https://doi.org/10.1002/gepi.22563