Causal Inference in Genetic Trio Studies

Stephen Bates[1], Matteo Sesia[1], Chiara Sabatti[1,2], and Emmanuel Candès[1,3]

[1]Department of Statistics, Stanford University
[2]Department of Biomedical Data Science, Stanford University
[3]Department of Mathematics, Stanford University

February 21, 2020

Abstract

We introduce a method to rigorously draw causal inferences—inferences immune to all possible confounding—from genetic data that include parents and offspring. Causal conclusions are possible with these data because the natural randomness in meiosis can be viewed as a high-dimensional randomized experiment. We make this observation actionable by developing a novel conditional independence test that identifies regions of the genome containing distinct causal variants. The proposed *Digital Twin Test* compares an observed offspring to carefully constructed synthetic offspring from the same parents in order to determine statistical significance, and it can leverage any black-box multivariate model and additional non-trio genetic data in order to increase power. Crucially, our inferences are based only on a well-established mathematical description of the rearrangement of genetic material during meiosis and make no assumptions about the relationship between the genotypes and phenotypes.

Keywords: trio, transmission disequilibrium test (TDT), family-based association test (FBAT), genome-wide association study (GWAS), causal discovery, false discovery rate (FDR), conditional independence testing

1 Introduction

Over the last 20 years, genome-wide association studies (GWAS) have established the association of thousands of genetic markers with phenotypes of interest, and these studies have the potential to shed light on essential scientific questions and to enable personalized medical care [1]. Analyzing GWAS data with increasingly large sample sizes poses new statistical challenges, however; in this new regime, any statistical association between a genetic variant and phenotype will now be detectable, including many irrelevant associations arising from non-genetic factors such as differing environmental conditions. While there are existing methods to mitigate this problem [2, 3, 4], such methods are not guaranteed to remove it entirely, so, with large sample sizes, they may leave many detectable associations that do not represent interesting biological activity. As such, a fundamental challenge for GWAS in the next decade is to move from detecting promising associations to rigorously establishing causality. On that premise, the most trustworthy way to ascertain that a statistical association is causal is to use a randomized experiment [5], and parent-offspring duo or trio data record such an experiment in the sense that the locations of the recombination points during meiosis are randomized by nature. Building on this, we propose a method that analyzes the placement of such sites to provably report only biologically meaningful regions of the genome.

1.1 Our contribution

In this work, we first articulate how causal inference is possible in the trio design and then introduce the *Digital Twin Test*: a method for finding causal regions that is immune to confounding variables. Our contribution has four components:

1. **Establishing causality in the trio design** We formalize the existing notion that family studies are immune to population structure, showing how to leverage the trio design to make causal inferences in a rigorous statistical sense. We formulate a broad class of tests for the null hypothesis of no causal effect, which includes existing methods such as the transmission disequilibrium test (TDT) [6, 7] and some of its variations for quantitative traits.

2. **A method for discovering causal regions** We introduce the Digital Twin Test to make causal inferences using trio data, improving on several aspects of the existing methods.

   (a) **Identifying distinct causal regions** Our method provably localizes causal variants within explicit windows along the genome, clearly showing the user when there are distinct causal effects. By contrast, although it is not widely known, the TDT is testing a less exact chromosome-wide null, so spurious findings arise from correlations among sites on the chromosome—see Section 4.1 for an example.

   (b) **Testing multiple hypotheses** Our method deals with multiple comparisons in a precise way, controlling either the family-wise error rate or the false discovery rate (FDR) without the need for a conservative Bonferroni correction. The heart of our solution is the creation of *independent* p-values for distinct regions of the genome, which can then be used together with more powerful multiple testing procedures.

   (c) **Leveraging black-box models and subject matter knowledge** Our method increases power by incorporating any multivariate model and subject matter information. Critically, the error rate guarantees of the method do not rely whatsoever on the correctness of the prior information or of the phenotype model.

While our inferences are based on trio data, our method can take advantage of additional case-control or population GWAS data to greatly increase power *while retaining the certified causal inferences*. Since trio samples are harder to collect than case-control or population samples, the techniques we describe do not replace existing methods for the latter, but instead work together with them to rigorously establish that the detected associations are due to causal variants. Lastly, we highlight that our approach is flexible and naturally applies to binary, quantitative, or time-to-onset response variables.

1.2 Trios, causality, and confounding in the literature

We begin by outlining related methods from the genetics literature. In inheritance, each parent transmits one complete set of DNA to the progeny, copying large continuous segments from that parent's two strands with occasional switches. Geneticists have long exploited the randomness in this process to identify meaningful associations [8, 9, 10]. The launching point for this work is the TDT [6, 7], which checks if a given allele is inherited more or less frequently than would be expected due to chance among the observations with a disease. If the transmission frequency deviates from the baseline frequency, the TDT reports an association. Beyond the original TDT,

additional techniques for more complex, partially observed pedigrees [11, 12, 13] and utilizing multiple markers [14, 15, 16] have been developed; these are known as *family-based association tests*. These methods are robust both to modeling assumptions about the relationship between the trait and the genotypes and to confounding due to *population structure*, namely, the presence of subpopulations with different allele frequencies [e.g., 17]. Moreover, these techniques can be extended to address quantitative traits [18, 19, 20], although, unlike the TDT, some of these methods require parametric assumptions about the relationship between the phenotypes and genotypes [21]. Similarly, the method of pseudocontrols [22, 23] computes the likelihood of an observed transmission pattern of a small number of alleles based on a posited model for the genotypes and phenotypes. To address the multiple comparisons issue arising from looking at several variants at once, [24] shows how to decouple the selection of promising markers from the final construction of a p-value from family-based association tests.

Turning to the statistics literature, causal inference is concerned with understanding what would happen if one manipulated the system under study, such as if one changed the SNPs in a small region of the genome while leaving all other conditions the same. This was first explicitly formulated in the context of randomized experiments by Fisher and Neyman [25, 26, 27], who pointed out that randomization removes the effect of *confounders*, unmeasured variables that create dependence between a covariate and the response even when directly changing the covariate would not affect the response. Therefore, any observed dependence in a randomized experiment must be due to a causal effect. This notion was later formalized for observational studies with the potential outcomes framework [5, 28] and, alternatively, with structural equation models [29, 30, 31]. While there is a spirited debate in the research community about when each is more fruitful, in this work we will demonstrate that our causal inferences are valid simultaneously in both frameworks.

Within causal inference, the problem of learning the structure of the true underlying model from data is known as *causal discovery* [e.g., 32, 33, 34]. General methods for causal discovery exist, although they typically require a large number of conditional independence tests and the assumption that such tests can be carried out without any statistical error [35, 36]. As a result, finite-sample results are rare. This work also uses conditional independence testing as the foundation for causal discovery but builds upon the conditional randomization test [37] to give finite-sample statistical guarantees. Our approach is also related to that of knockoffs [38, 37] which also provides finite-sample statistical guarantees and has been successfully deployed to analyze GWAS data [39, 40, 41], although the connection with causal discovery was not previously developed.

Lastly, we pause to clarify the relationship of this work with Mendelian randomization and fine-mapping. First, Mendelian randomization [42, 43] is a technique that uses a pre-specified genetic variant as an instrumental variable in order to determine the magnitude of the causal effect of an exposure (e.g., smoking) on an outcome (e.g., lung cancer). By contrast, here we are instead interested in identifying which of many genomic regions have a causal effect on an outcome. Second, methods for *fine-mapping* seek to reduce a large region of GWAS detections to a smaller set that is expected to contain the causal variant driving the association [e.g., 44]. There, the term "causal variant" indicates that one is operating at a high-resolution, whereas the current work is concerned with establishing that associations are truly biological and are not created by confounding.

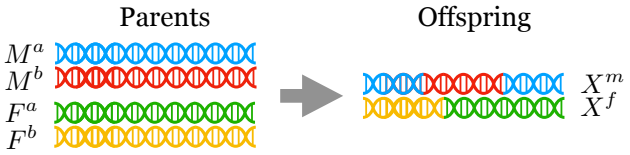

Figure 1: A visualization of the process of recombination on a single chromosome.

2 Causal inference in the trio design

2.1 Setting

Human cells have 46 DNA strands organized into 23 chromosome pairs; in each pair, one complete strand is inherited from the mother and one is from the father. In this work, we consider the case where we measure *single-nucleotide polymorphisms* (SNPs), sites on the genome where two possible alleles occur in the population, encoded as 0 or 1. The set of observed alleles for one entire strand is known as a *haplotype*. We consider the case where the haplotypes of n subjects and their biological parents at p sites are known, denoted as follows:

$$\begin{array}{rcc} \text{subjects' haplotypes} & (X_1^m,\ldots,X_p^m)\in\{0,1\}^{n\times p}, & (X_1^f,\ldots,X_p^f)\in\{0,1\}^{n\times p};\\ \text{mothers' haplotypes} & (M_1^a,\ldots,M_p^a)\in\{0,1\}^{n\times p}, & (M_1^b,\ldots,M_p^b)\in\{0,1\}^{n\times p};\\ \text{fathers' haplotypes} & (F_1^a,\ldots,F_p^a)\in\{0,1\}^{n\times p}, & (F_1^b,\ldots,F_p^b)\in\{0,1\}^{n\times p}. \end{array}$$

For convenience we denote the matrix of offspring *genotypes* as $X = X^m + X^f$, the j-th column of X as X_j, the i-th row of X as $X^{(i)}$ (this is the i-th individual), and the set of all ancestral haplotypes as $A = (M^a, M^b, F^a, F^b)$.

Our method takes the haplotypes as given, even though typically only the genotypes are directly measured in a GWAS study. Haplotypes are then reconstructed algorithmically through *phasing* [45]. While experimental techniques are being developed to directly measure haplotypes, these are not yet widespread. We instead take the phased haplotypes as a reasonable approximation, since phasing is known to be accurate with family data [45, 46, 47]. In a simulation with a synthetic population with known ground-truth haplotypes, we find that our method performs identically with known haplotypes and computationally phased haplotypes; see Appendix G.3.

Crucially, the distribution of the offspring genotypes X conditional on the parental haplotypes A is known [48]. Informally, the model for a single offspring is this: for the haplotype X^m inherited from the mother, the SNP X_j^m is inherited either from M_j^a or from M_j^b, with equal probability. Furthermore, long continuous blocks of X^m are jointly inherited either from M^a or M^b, with occasional switches at *recombination* sites; see Figure 1 for an illustration and Section 2.5 for a formal description of this process. Throughout, we will leverage our knowledge of the recombination mechanism in order to carry out hypothesis tests.

2.2 Establishing causality

We now demonstrate that it is possible to draw causal inference from trio data by formulating the inheritance process as a high-dimensional randomized experiment. The main idea is to condition on the parental haplotypes: once these are fixed the remaining randomness in meiosis is due to cell-

level biological processes that are independent of possible confounders, so the resulting inferences are immune to these factors.

We begin with a concrete example of confounding in GWAS. Suppose we have a genetic study involving two populations, such as Europeans and East Asians, and we wish to study whether a SNP X_j affects the cholesterol level Y. Next, suppose that the distribution of cholesterol levels differs in the two populations, and furthermore that the distribution of SNP X_j also differs. As a result, there is a valid statistical association between X_j and Y, but this statistical association may or may not represent a causal effect. That is, if we manipulated the SNP X_j, it may or may not change the cholesterol levels of the subjects. The association could instead be the result a confounder; for example, if the consumption of dairy products leads to higher cholesterol and Europeans consume more dairy products on average, then the SNP X_j will be associated with Y, even when it has no causal effect. Since randomized experiments only detect causal effects, to circumvent the above problem we could, in principle, flip a fair coin, set the value of X_j accordingly, and then check for an association with Y. While we of course do not carry out such experiments on people, we can exploit a similar experiment occurring in nature.

Formally, let $B \mid C$ denote the distribution of a random variable B given the observed value of a random variable C, and let $B \perp\!\!\!\perp C \mid D$ denote that B is conditionally independent of C given D. Consider testing the null hypothesis that a SNP X_j is independent of the response Y after conditioning on the parental haplotypes:

$$H_0 : X_j \perp\!\!\!\perp Y \mid A. \tag{1}$$

Next, let Z be a potential confounder, such as dairy consumption above. The critical observation is that for essentially all possible confounders of concern in genetic studies, the distribution of the genotype at site j given the parental haplotypes does not change given knowledge of Z, since the randomness in inheritance is a result only of random biological processes independent of Z. To make this precise, we define the following set of possible confounders.

Definition 1 (External confounder)**.** *We say that a random variable Z is an* external confounder *if the distribution of the offspring's haplotypes given the parental haplotypes does not change given knowledge Z:*

$$X \mid (A, Z = z) \quad \stackrel{d}{=} \quad X \mid (A, Z = z') \quad \text{for any } z \text{ and } z'. \tag{2}$$

The relation in (2) is true for the offspring's dairy consumption, for example, as well as for all environmental conditions occurring after conception. This relationship implies that Z is independent of the offspring's SNP j given the parental haplotypes:

$$Z \perp\!\!\!\perp X_j \mid A.$$

This then implies that if there is an association between Y and X_j after conditioning on the parental haplotypes, then the association is not due to the confounder Z:

$$Y \not\perp\!\!\!\perp X_j \mid A \implies Y \not\perp\!\!\!\perp X_j \mid (A, Z).$$

Returning to the language of hypothesis testing, this proves that if we test the null hypothesis in (1), we automatically account for the random variable Z. We record this fact formally:

Theorem 1 (Conditioning on parents accounts for external confounders)**.** *Let Z be an external confounder, i.e., Z satisfies the relation in* (2)*. Then, any valid test of the null hypothesis in* (1) *is also a valid test of the stronger null hypothesis*

$$H_0' : Y \perp\!\!\!\perp X_j \mid (A, Z) \tag{3}$$

that accounts for the confounder Z.

In words, if we test the hypothesis in (1), which is possible based on observed trio data, then we have perfectly adjusted for the confounder Z, *even if it is not specified or measured in the data.* Thus, if we reject the null in (1), it cannot be the case that X_j and Y are dependent due to an external confounder Z. We emphasize that rejecting a test of the null hypothesis in (1) does not yet imply that X_j itself is the causal SNP, since the dependence between X_j and Y may be the result of a causal SNP elsewhere on the chromosome, but it does imply that there is an association on the chromosome that is not the result of external confounding. We will show how to explicitly localize the causal regions later in Section 4.

2.3 Connection to structural equation modeling

Thus far, we have made statements about conditional independence relationships without explicitly using existing formalisms for causal inference. Here, we formulate our results as a structural equation model to make the connection with the existing literature explicit, and we similarly formulate our results in the potential outcomes framework in Appendix D.

Consider a structural equation model involving the variables A, X, Y and the external confounder Z. For a response Y, we assume that X can only cause Y and not the reverse, which is reasonable because a subject's genotype is fixed after conception. We further know that the parental haplotypes A cause X and not the reverse. We also assume that Z causes Y since the reverse case does not result in confounding. Finally, by definition the external confounder Z is conditionally independent of X given A, which implies that there is no causal effect from X to Z. The corresponding structural equation model is

$$(A, Z) = f_{AZ}(N_{AZ}), \qquad X = f_X(A, N_X), \qquad Y = f_Y(X, Z, N_Y),$$

where f_{AZ}, f_X, and f_Y are fixed functions and N_{AZ}, N_X and N_Y are independent uniform $[0, 1]$ random variables; see Figure 2 for a graphical representation. Within this model, rejecting the hypothesis in (1) implies that there is a causal effect from X to Y (although not necessarily from the specific SNP X_j to Y). In this sense, testing the hypothesis in (1) is a formal causal inference. Crucially, our proposed method will test this hypothesis without the need to specify or restrict f_{AZ} or f_Y.

2.4 Discussion of possible confounders

Virtually all confounders of concern in genetic studies do not affect the transmission of the genetic information from parents to offspring, and are thus external confounders which are correctly accounted for in the trio design by Theorem 1. We list the most important examples below.

- **Environmental conditions after conception** The mechanism for producing X from A is unaffected by anything occurring after conception.
- **Population structure, ethnic composition, and geographic location** The mechanism for producing X from A does not change with subpopulation information, ethnicity, or geographic location.
- **Cryptic relatedness** The presence of distantly related individuals in a sample does not change the distribution of X given A, even if this relatedness is unknown and unspecified.

(a) (b)

Figure 2: A graphical depiction of the causal argument in Section 2.2. Panel (a) shows that the random variable Z can create an association between X_j and Y, even if there is no causal effect. Panel (b) shows that conditional on the parental haplotypes A, the external confounder Z is independent of the offspring's genotype X_j. As a result, Z cannot be responsible for the remaining association between the genotype X_j and the trait Y.

- **Family effects, altruistic genes** Information about the quality of the environment caused by parental behavior does not impact the distribution of X given A.
- **Assortive mating** Tests of (1) condition on the observed mating pattern, making them immune to this form of confounding.

By contrast, the following are not external confounders:

- **Germline mutations** A few environmental factors of the parents can affect the inheritance process, such as the exposure of a parent to radiation, which changes the distribution of the offspring by increasing the frequency of mutations. While this does affect the model for inheritance in principle, we do not expect this to practically invalidate tests of the null in (1). In any case, this is a narrow set of possible confounders.
- **Unmeasured SNPs** In typical studies, only a subset of SNPs are sequenced. Knowledge of a subject's unmeasured SNPs gives additional information about the distribution of X given A. Later in this work, we will show how to localize the unmeasured causal SNPs.

Soon, in Section 3.1, we develop a family of tests of the null in (1), but we now pause for two additional remarks about our inferences. First, the reader may wonder how mutations that occur later in a subject's life, such as mutations caused by sun or radiation exposure, fit with the above exposition. Such mutations originate in a single cell, and only descendant cells carry the mutation. As a result, any mutations not present at the earliest stages of development will only be present in a very small minority of a subject's cells. Genotyping techniques use the consensus of measurements from a large number of cells, and so they essentially measure the genotype at conception.

Second, there is potential for selection bias in all genetic studies, since some individuals are more likely to be included in a sample than others. Tests of the null in (1) remove the effect of selection bias, except for that directly caused by the offspring's genetic composition. Any potential selection bias due to external confounders, such as geographic location, is automatically accounted for by Theorem 1. However, if a SNP X_j causally influences the probability of inclusion in a study, then such a SNP is not null according to our null hypothesis in (1), so it may be detected by such a test.

2.5 The randomness in inheritance

In preparation for a test of hypothesis in (1), we now formally describe the distribution of offspring's genotype given the parental haplotypes. The biological mechanism governing inheritance is well-understood, and this will serve as the backbone of our statistical inference. In particular, the process by which a subject's two haplotypes arise from the parental haplotypes was formalized by Haldane as a hidden Markov model (HMM) [48]. Without loss of generality, we describe the model for a single observation on one chromosome; in the general case, each chromosome of each observation is an independent instance of this model. For concreteness, we focus on X^m.

Latent Markov chain Let the random vector $U^m \in \{a, b\}^p$ indicate the following:

$$U_j^m = \begin{cases} a & \text{if site } j \text{ is copied from the mother's 'a' haplotype,} \\ b & \text{if site } j \text{ is copied from the mother's 'b' haplotype.} \end{cases}$$

Our model is that U^m is distributed as a Markov chain, where

$$P(U_1^m = a) = \frac{1}{2},$$

and

$$\begin{aligned} P(U_j^m = u_{j-1}^m \mid U_{1:(j-1)}^m = u_{1:(j-1)}^m) &= P(\text{even } \overbrace{\#\text{ of recombinations between } j-1 \text{ and } j}^{\text{modeled as Poisson with mean } d_j}) \\ &= \frac{1}{2}(1 + e^{-2d_j}). \end{aligned}$$

Here, d_j is the *genetic distance* between SNPs $j-1$ and j, which is fixed and known. Note that the genetic distance is not always proportional to the physical distance due to recombination hotspots: regions that have more frequent recombination events [49, 50].

Emission distribution Conditional on U^m, each X_j^m is independently sampled from

$$P(X_j^m = M_j^{(u_j^m)} \mid U_j^m = u_j^m) = 1 - \epsilon.$$

Here, ϵ is the probability of a de novo mutation, which for humans is about $1 \cdot 10^{-8}$ [51]. The analogous HMM describes the distribution of X^f given F^a and F^b, which is taken to be independent of X^m given A.

3 Testing the global causal null

The remainder of this work develops randomization tests to address two questions of interest. First, we wish to determine whether a trait has any genetic basis, that is, to test the global null. Second, we wish to find regions of the genome that contain causal variants. To clarify the relationship between these questions and our choice of null hypothesis, consider H_0 in (1). This hypothesis probes the association between the trait of interest and SNP X_j, without considering the role of

other genetic variants. It is non-null either when X_j is directly related to the trait or when X_j is merely associated with a causal genetic variant. Therefore, to establish that X_j itself is the causal variant, we additionally need to ensure that the association between X_j and Y is not due to any other genetic variables. One removes these possible confounders by conditioning on them, testing instead hypotheses that check the dependence of Y and X_j after accounting for the rest of the genome [37, 41]. We will describe how to do this in detail soon, but we start by noting that once we condition on the parental haplotypes, the SNPs on different chromosomes are independent. As a result, in the trio design even global null tests can localize signals to a single chromosome.

3.1 The full-chromosome Digital Twin Test

We begin by testing the a full-chromosome global null. Let $\mathcal{C} \subset \{1, \ldots, p\}$ be the set of SNPs on the given chromosome. We now consider the hypothesis

$$H_0 : X_{\mathcal{C}} \perp\!\!\!\perp Y \mid A, \tag{4}$$

where $X_{\mathcal{C}}$ denotes $(X_j)_{j \in \mathcal{C}}$. The *Digital Twin Test* of the null hypothesis in (4)—presented in Algorithm 1—compares an observed offspring to synthetic offspring from the same parents (the "digital twins"; see Figure 3) using an arbitrary measure of feature importance or predictive accuracy.

Algorithm 1: The Digital Twin Test

— Compute the test statistic on the true data:

$$t^* = T((X_{-\mathcal{C}}, X_{\mathcal{C}}), Y).$$

for $k = 1, \ldots, K$ **do**

— Sample the digital twins $\widetilde{X}^m_{\mathcal{C}}$ and $\widetilde{X}^f_{\mathcal{C}}$ from the distribution from Section 2.5, independent of Y (see Appendix C for an explicit sampler).

— Compute the test statistic using the digital twins:

$$t_k = T((X_{-\mathcal{C}}, \widetilde{X}^m_{\mathcal{C}} + \widetilde{X}^f_{\mathcal{C}}), Y).$$

end

— Compute the quantile of the true statistic t^* among the digital twin statistics $t_1, \ldots, t_K$:

$$v = \frac{1 + \#\{k : t^* \leq t_k\}}{K + 1}.$$

Result: A p-value v for the conditional hypothesis in (4).

In Algorithm 1, $X_{-\mathcal{C}}$ denotes $(X_j)_{j \notin \mathcal{C}}$, and $T(\cdot)$ can be any statistic. This algorithm returns a p-value, and the corresponding level α hypothesis test rejects when this p-value is less than α. We record formally that this test accounts for external confounders; see Appendix B for all proofs.

Proposition 1 (The Digital Twin Test is immune to external confounders)**.** *The Digital Twin Test is a valid test of the hypothesis*

$$H_0^{\mathcal{C}} : Y \perp\!\!\!\perp X_{\mathcal{C}} \mid (A, Z), \tag{5}$$

where Z is any (possibly unknown) external confounder. That is, if the null hypothesis in (5) *holds, then for all $\alpha \in (0, 1)$, the output v of Algorithm 1 obeys*

$$P(v \leq \alpha) \leq \alpha.$$

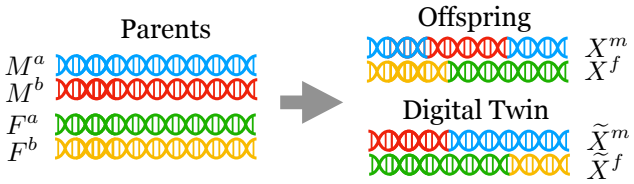

Figure 3: A visualization of a digital twin.

The Digital Twin Test is a natural randomization test in a trio design because it replicates the random mechanism generating the data. By contrast, a traditional permutation test can only test for a marginal association between X_j and Y, which does not lead to valid causal inferences; see Appendix F.

Many existing tests fall into this family. Notably, the TDT is a special case of the Digital Twin Test with the test statistic

$$T^{(\mathrm{TDT})}(X) = \sum_{i=1}^{n} X_j^{(\mathrm{i})} \mathbb{I}_{\{Y_i=1\}}, \tag{6}$$

where $\mathbb{I}_B$ denotes the indicator of event B, except that the Digital Twin Test uses an exact, finite-sample rejection threshold, whereas the TDT uses an asymptotic approximation. Similarly, the quantitative TDTs [19] and [20] are also special cases of the above procedure. Moreover, the Digital Twin Test can exploit arbitrary black-box machine learning models, such as deep neural networks, gradient boosting, random forests, and penalized regression to form a test statistic $T(\cdot)$ that incorporates information from multiple sites in a data-driven way; see (7) below for a concrete example. More sophisticated models can explain away more of the variation in the phenotype, leading to more sensitive tests. Thus we can see that the Digital Twin Test framework unifies many existing procedures while incorporating varying disease models, subject matter knowledge, fitting algorithms, principal component corrections, screening and replication, and so on, without requiring a new mathematical analysis for each case. While well-chosen models will lead to more powerful tests, we emphasize that the validity of the automatic, finite-sample inference does not depend on the correctness of the chosen model.

3.2 Incorporating external GWAS data

The Digital Twin Test can also leverage large external GWAS data sets that do not contain trio observations to increase power. This is important because such data sets are common and will typically have larger sample sizes than trio data sets. To this end, we can use the external GWAS to find a powerful test statistic $T(\cdot)$, as suggested by [52]. For example, suppose we fit a penalized linear or logistic regression model on the external GWAS data to obtain an estimated coefficient vector $\hat{\beta}$. Then, on the trio data, we can use the Digital Twin Test with test statistic

$$T(X,Y) = \begin{cases} -\sum_{i=1}^{n} (\hat{\beta}^\top X^{(\mathrm{i})} - Y_i)^2 & \text{for real-valued } Y, \\ -\sum_{i=1}^{n} -Y_i \log(\frac{e^{\hat{\beta}^\top X^{(\mathrm{i})}}}{1+e^{\hat{\beta}^\top X^{(\mathrm{i})}}}) - (1-Y_i)\log(\frac{1}{1+e^{\hat{\beta}^\top X^{(\mathrm{i})}}}) & \text{for binary } Y \end{cases} \tag{7}$$

(the negative squared loss and the negative logistic loss). In words, the Digital Twin Test with this test statistics is asking, "are the residuals smaller when I use the real genotypes to predict the response, compared to when I use the digital twin genotypes to predict the response?" If the residuals are systematically smaller, it must be because of a causal effect, and the Digital Twin Test rejects the null hypothesis. We will explore this approach in simulation in Section 5.

3.3 Parent-offspring duos and other pedigrees

The Digital Twin Test can also be applied to offspring for whom only one parent is genotyped, with a small adjustment. The modification is simple: whenever a parent is unknown the algorithm fixes the offspring's haplotype from that parent. For example, if F^a and F^b are unknown then in Algorithm 1 we set $\widetilde{X}_g^f = X_g^f$ in each iteration of the loop. An analogous version of Theorem 1 continues to hold in this setting, and so we still detect only causal regions.[1]

Furthermore, the Digital Twin Test can be applied to data with a variety of pedigree structures. One can select any set of duos or trios from the pedigree, with the restriction that no offspring in a duo or trio is an ancestor of any other offspring in a duo or trio. Even though individuals may be related, the locations of their recombination sites are independent. Moreover, only the parent-offspring relations need to be known, not the full pedigree.

4 Localizing causal regions

4.1 Linkage disequilibrium in the trio design

SNPs on the same chromosome are dependent, a phenomenon referred to as *linkage disequilibrium* (LD), so the Digital Twin Test and TDT can only rigorously test the full-chromosome causal null hypotheses discussed in Section 3. Before presenting in this section a version of the Digital Twin Test to localize causal SNPs, we highlight the current limitation with a practical example.

We create a synthetic population of 2,500 second-generation admixed individuals whose parents are the children of one ethnically British individual and one ethnically African individual. The haplotypes of the parents are real in the sense that they are phased haplotypes from the UK Biobank data set [53]. For simplicity, we create a binary synthetic response Y from a logistic regression model with a single causal SNP. We choose a signal strength such that that the heritability of the trait is 18% and an intercept such that 20% of the observations have $Y = 1$. The causal SNP is chosen at a site with a large difference in allele frequency between the British and African populations. Then, we carry out the TDT at each site and report the p-values in a Manhattan plot in Figure 4. Note that even if we demand that the p-values are smaller than the genome-wide significance threshold of $5 \cdot 10^{-8}$, the TDT reports discoveries all across the chromosome. We compare this to an identical simulation comprised only of British individuals in the center panel of Figure 4. Here, the TDT only reports discoveries near the true causal SNP.

The TDT behaves differently in these two populations because of the different correlation structure after conditioning on the parental haplotypes. In the admixed population, there are large correlations between sites far away, but not in the British population; see the right panel of Figure 4. Because of the large LD in the admixed population, testing the full-chromosome causal null in (1) cannot give reliable information about the location of the causal SNP—this limitation applies both to the TDT and to the Digital Twin Test from Section 3. This weakness of the TDT has been noticed before in in the admixed setting [54]. That work developed an analytical correction based on population-genetic quantities. Instead, we will next extend the Digital Twin Test to provably localize the causal regions using conditional independence testing.

[1] We simply condition on X^f rather than F^a and F^b for each unit where the father's haplotypes are unknown, and so on.

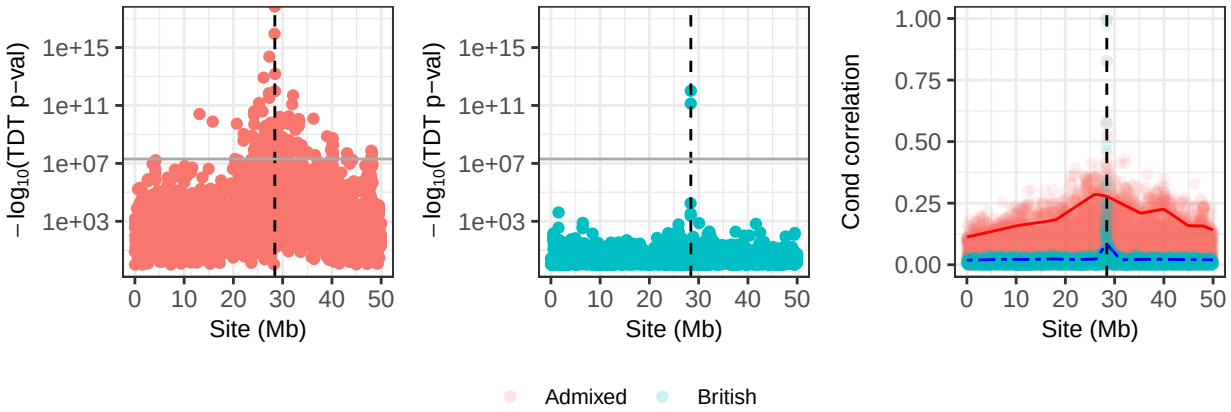

Figure 4: Results of the TDT in two populations. Left and center: Manhattan plots on chromosome 22, which contains the one true causal SNP, indicated with a dashed vertical line. The genome-wide significance threshold is shown with a gray horizontal line. The left panel shows an admixed population, whereas the center panel shows a British population. Right: a plot of the absolute correlations between the causal SNP and the other SNPs, conditional on the parental haplotypes. The red solid and blue dot-dashed curves indicate a smoothed 90% quantile of the absolute correlation with the causal SNP across the chromosome, for the admixed and British populations, respectively.

4.2 The Local Digital Twin Test

Since the location of the causal SNPs cannot be reliably found from the TDT or any other full-chromosome test, we next introduce the *Local Digital Twin Test* to identify distinct causal regions. Our method partitions the genome into disjoint regions and then constructs a p-values for the hypothesis that the region contains no causal SNPs. Because we investigate many regions simultaneously, it is critical to leverage modern methods for multiple testing, such as the Benjamini–Hochberg procedure [55], in order to have higher power than conservative methods like Bonferroni. To this end, we will also develop a technical modification of the Local Digital Twin Test that yields *independent* p-values, which can be used as inputs to more powerful multiple testing procedures.

Formally, let G be a partition of $\{1, \ldots, p\}$. For each group of SNPs $g \in G$, we will consider the hypothesis

$$H_0^g : Y \perp\!\!\!\perp X_g \mid (X_{-g}^m, X_{-g}^f, A). \tag{8}$$

In words, this is the hypothesis that knowing the SNPs in group g is not informative about the response once we know the remaining SNPs and the parental haplotypes. This ensures that any detections are the results of causal SNPs in the region g rather than elsewhere on the chromosome. In what follows, we will consider groups of SNPs g that form a continuous block of the chromosome, but this is not required. Notice that larger group sizes correspond to weaker statistical statements and hence the corresponding tests have higher power [e.g., 56, 41].

The Local Digital Twin Test—presented in Algorithm 2—tests the null hypothesis in (8) by again creating synthetic offspring from a subject's parents (the "local digital twins"). Unlike before, the synthetic offspring are now additionally constrained to match outside the region g. That is,

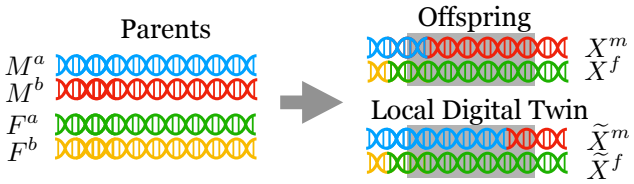

Figure 5: A visualization of a local digital twin. The gray shaded region represents the group g; the local digital twin always matches the true offspring outside this region.

they are sampled from the distribution of

$$(X_g^m, X_g^f) \mid (X_{-g}^m, X_{-g}^f, A). \tag{9}$$

See Figure 5 for an illustration.

Algorithm 2: The Local Digital Twin Test

— Compute the test statistic on the true data:

$$t^* = T((X_{-g}, X_g), Y).$$

for $k = 1, \ldots, K$ **do**

— Sample the digital twins $(\widetilde{X}_g^m, \widetilde{X}_f^m)$ from the distribution (9), independent from (X_g^m, X_g^f) and Y (see Appendix C for an explicit sampler).

— Compute the test statistic using the digital twins:

$$t_k = T((X_{-g}, \widetilde{X}_g^m + \widetilde{X}_g^f), Y).$$

end

— Compute the quantile of the true statistic t^* among the digital twin statistics $t_1, \ldots, t_K$:

$$v = \frac{1 + \#\{k : t^* \leq t_k\}}{K + 1}.$$

Result: A p-value v for the conditional hypothesis in (1).

This test both accounts for external confounders and localizes the causal regions, as stated next.

Proposition 2 (The Local Digital Twin Test is immune to external confounders)**.** *The Local Digital Twin Test is a valid test of the hypothesis*

$$H_0^g : Y \perp\!\!\!\perp X_g \mid (X_{-g}^m, X_{-g}^f, A, Z), \tag{10}$$

where Z is any (possibly unknown) external confounder. That is, under (10)*, for all $\alpha \in (0, 1)$, the output v of Algorithm 2 satisfies*

$$P(v \leq \alpha) \leq \alpha.$$

The Local Digital Twin Test p-values above are not guaranteed to be independent for null groups, so we present a modification that gives provably independent p-values in Appendix A; see

Algorithm 3 therein. The main idea is to isolate the randomness used to test each region. The details require substantial additional notation, however, so they are deferred to the Appendix.

Theorem 2 (Independence of null p-values)**.** *Algorithm 3 (defined in Appendix A) produces p-values v_g satisfying the following properties:*

1. *For groups g satisfying the Local Digital Twin Test null hypothesis in* (10)*, the distribution of v_g stochastically dominates the uniform distribution, i.e., v_g is a valid p-value.*
2. *The p-values for null groups are jointly independent.*

As before, any score function $T(\cdot)$ leads to valid inference, but some functions will lead to higher power. One option requiring only trio data is the statistic in (6). This gives a TDT-like test, with additionally localization guarantees at the price of lower power. A better option is to incorporate external GWAS data with the previously-discussed test statistic in (7).

4.3 Weighted and ordered testing

To further increase power, the external GWAS data should be used to prioritize the most promising regions of the genome for confirmatory testing with the Local Digital Twin Test; [57] gives a general discussion of incorporating weighted testing in GWAS and [58] show how to use information to improve the ordering for knockoff testing. As a concrete example, when using the test statistic in (7) one could order the hypothesis by decreasing value of

$$w_g = \sum_{j \in g} |\hat{\beta}_j|, \tag{11}$$

and then use the Selective SeqStep procedure [38] or an accumulation test [59] to give a final selection set with guaranteed FDR control. We numerically explore this approach in Section 5.

4.4 Resolution of the Local Digital Twin Test

The Local Digital Twin Test will depend primarily on those samples where a recombination is observed in the region g under study. For a single subject, consider $\widetilde{X}_g^m$ in Algorithm 2. Suppose that no recombination event occurs in the region g for the maternal haplotype X^m. Then, the subject's haplotype X^m is copied from the same of the mother's two haplotypes (for example, M^a) at both edges of g, so the most likely sample from distribution in (9) is such that the entire region including g matches M^a, in which case $\widetilde{X}_g^m = X_g^m = M_g^a$; see $\widetilde{X}^f$ and X^f in Figure 5. These samples are not informative, since there is no contrast between the true observations and the digital twins to use for inference. On the other hand, if a recombination takes place in the group g, then $\widetilde{X}_g^m$ and X_g^m will differ; see $\widetilde{X}^m$ and X^m in Figure 5. Putting this together, only samples that contain a recombination event in the group g will be of use.

This behavior limits the resolution of the Local Digital Twin test. In humans, the probability of a recombination event in a 1 Mb region is about a 1% [49]. Thus, for a data set of $100,000$ trios, each region of size 1 Mb would only have about $2,000$ samples with recombination events:

$$\underbrace{100,000}_{\text{sample size}} \times \underbrace{1\text{ Mb}}_{\text{size of region}} \times \underbrace{0.01\ \text{ recombinations / Mb}}_{\text{recombination rate}} \times \underbrace{2}_{\text{Haplotypes per subject}} = \underbrace{2,000.}_{\text{recombination events}}$$

As a result, the Local Digital Twin Test would essentially compare the $2,000$ subjects with recombination events to their digital twins. The remaining observations could be used for fitting the multivariate model, but would not be used for confirmatory inference. This reduction in effective sample size at fine resolution limits the precision with which the Local Digital Twin Test can make discoveries, and is the price we pay for making stronger causal inferences. Nonetheless, our goal is not single-SNP precision, but rather is to rigorously establish that there are distinct causal regions across the genome.

5 Simulation experiments

In this section, we examine the performance of the Digital Twin Test in semi-synthetic examples, focusing on the binary response case, so that the standard TDT can serve as a benchmark. We form our parent-offspring population by taking real haplotypes from the UK Biobank data set and sample offspring according to the recombination model from Section 2.5. In each experiment, we sample the offspring once and then repeat the generation of the synthetic phenotype multiple times, indicating the standard error with error bars. When presenting the results, we index the signal strength by *heritability*: a $[0, 1]$-valued scale defined in Appendix G.1. An R package implementing the methods below together with notebook tutorials is available at `https://github.com/stephenbates19/digitaltwins`.

5.1 Testing the full-chromosome causal null

We first examine the ability of the Digital Twin Test to test the full-chromosome causal null. In this simulation, we test only one hypothesis, so we seek to control the usual type-I error rate at the $\alpha = 0.05$ level. We create a synthetic population of $n = 2,500$ parent-child trios and generate a binary valued response coming from a sparse logistic regression model

$$\log\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = \beta_0 + \beta^\top X^{(\mathrm{i})}, \tag{12}$$

with 10 nonzero entries of β of equal strength, chosen uniformly at random. The intercept β_0 is chosen so that the fraction of cases is $50\%, 20\%$, or 5%. We use $p = 6,820$ SNPs from chromosome 20, which has width 63 Mb. We emphasize that the above gives a well-defined structural equation model on (X, Y), and the nonzero entries of the coefficient vector β correspond to the SNPs that have a causal effect.[2]

In order to use the Digital Twin Test for confirmatory analysis, we take an external GWAS data set that contains $7,500$ other non-trio observations from the UK Biobank. With these data, we fit a ℓ_1-penalized logistic regression model (with regularization parameter chosen by cross-validation) to obtain a predictive model for the trait. We denote the resulting coefficient estimate $\hat{\beta}$. Then, we apply the Digital Twin Test on the trio data with feature importance statistic in (7) to produce a single p-value, rejecting when it falls below $\alpha = 0.05$.

We take the TDT as a natural benchmark, interpreting its output in two alternative ways. First, we take the minimum p-value after applying the TDT at every SNP and then Bonferroni

[2]The reader may wonder about the identifiability of this model. Note that both X and Y are random variables, so provided that in the distribution of $X^{(\mathrm{i}),m} + X^{(\mathrm{i}),f}$ is not contained in a subspace of rank less than p, then this model is identified.

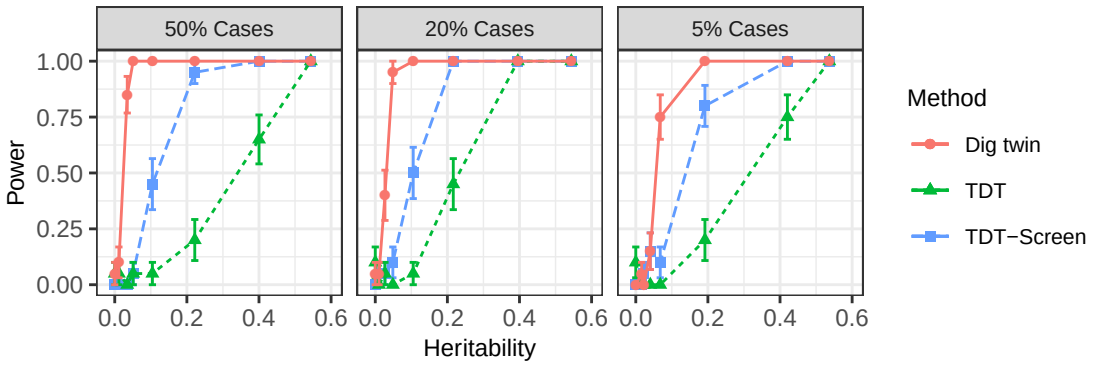

Figure 6: Power of the Digital Twin Test compared to TDT benchmarks for testing the full-chromosome causal null.

correct it (this does not use the non-trio GWAS data). Second, we compute the Bonferroni-adjusted minimum p-value only on the coordinates with nonzero coefficients in the lasso fit $\hat{\beta}$ on the external GWAS data. Because $\hat{\beta}$ is sparse, this method has a less severe Bonferroni correction and may be more powerful than the other TDT procedure.

Since all three methods are valid tests of the null hypothesis that there is no causal SNP on the chromosome, we directly compare their power in Figure 6, using 20 independent realizations for each data point. We find that the Digital Twin Test has higher power than the TDT, even when the latter attempts to leverage the external GWAS as a screening step. The leftmost point in each panel is the null case with zero heritability; the empirical error of the Digital Twin Test does not exceed the nominal level of $\alpha = 0.05$ in any of the three cases.

5.2 Localization

We now examine the ability of the Local Digital Twin Test to identify causal regions. Here, we use $p = 591,513$ SNPs on chromosomes $1 - 22$, split into 532 pre-determined groups of size approximately 5 Mb. The response is again generated from the logistic regression model in (12), and the number of non-zero coefficients in the true causal model is varied as a control parameter. We consider a sample of $n = 10,000$ trios with an external GWAS of size $50,000$ used to fit a logistic regression model $\hat{\beta}$ as in the previous section. The fitted coefficients $\hat{\beta}$ are used to form the test statistic in (7). Here, we take the nominal level for the FDR to be $\alpha = 0.2$. Each experiment is repeated 10 times. Additional technical details about these simulations can be found in Appendix G.2.

We compare the following procedures:

1. **Digital Twin Test—Accum** We apply the Local Digital Twin Test at each of the groups to obtain one p-value per group. We also use the external GWAS data to order the regions from most to least promising as in (11), and use an accumulation test [59] to produce a final set of discoveries. This method is guaranteed to control the FDR.[3]

[3]Strictly speaking, this procedure controls a modified version of the FDR (see [59]), but the difference will be unimportant in settings with a large number of discoveries. This is a property of the accumulation test, not the Local Digital Twin Test, and other procedures can be used for standard FDR control.

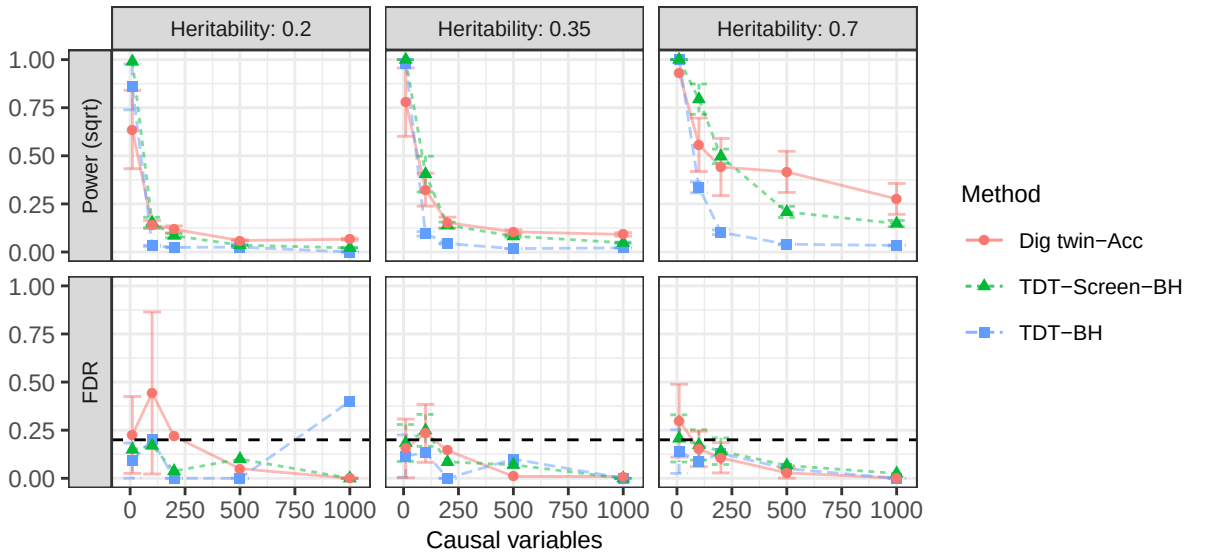

Figure 7: Performance of the Local Digital Twin Test and TDT in the binary-response full-genome simulations from Section 5.2. Here, error bars give one standard deviation, and the dashed horizontal line indicates the nominal FDR level

2. **TDT—Sreen—BH**: For each group, we apply the TDT to the SNPs with nonzero entries of $\hat{\beta}$, the model fit on the external GWAS. Then, we report the minimum p-value after adjusting it with Bonferroni. Lastly, we apply the Benjamini–Hochberg procedure to report a set of groups. This method assumes the TDT p-values are valid for the group null hypothesis in (8), which is not fully correct, so this method does not have formal guarantees.

3. **TDT—BH**: We proceed as above, except that we apply the TDT to all SNPs. This method also incorrectly assumes the TDT p-values are valid for the group null hypothesis in (8), so it does not have formal guarantees.

We report the results in Figure 7. The Local Digital Twin Test has generally comparable power to the screened TDT method with Benjamini–Hochberg, with moderate power improvements for traits caused by many SNPs. We report on a similar experiment with a continuous response in Appendix G.4.

Although the TDT benchmarks empirically control the FDR here, we emphasize that it formally tests the full-chromosome null in (1), so it is not valid for our goal of localizing the causal SNPs into the given groups. The group hypothesis in (8) can only be rigorously tested by a test of the null in (1) if the SNPs in different groups are independent. If the TDT-based group p-values were valid, these two benchmarks would control the FDR. This may be a reasonable approximation within this experiment, because the groups are wide and the population is homogeneous, so the LD decays rapidly. However, this assumption fails in other cases, as we previewed in Section 4.1 and will return to next.

5.3 Spurious discoveries with the TDT

We have seen in Section 4.1 an example where the TDT makes false discoveries throughout the chromosome because it does not account for LD. Here, we revisit this example more carefully,

demonstrating that the TDT-based benchmarks from Section 5.2 can dramatically fail to control type-I errors for the group hypotheses.

For simplicity, we perform the TDT at each SNP and report a SNP as significant if the p-value is below the genome-wide significance level, $5 \cdot 10^{-8}$. We use this demanding threshold to emphasize that even the most conservative existing methods violate type-I error control here. Indeed, recall that in the left-hand panel of Figure 4, we find rejections all across the chromosome, even though there is only one causal SNP. If the heritability is large, Figure 8 shows that spurious associations at least 1 Mb away from the causal SNP occur with high probability; here, each experiment is repeated 20 times. Moreover, spurious associations occur as far away as 15 Mb from the causal SNP, reinforcing that one should not interpret the TDT p-values as providing evidence of distinct causal findings.

Proceeding as in Section 5.2, we split the chromosome into 25 groups of size 2 Mb and consider the procedure that reports a group as significant if any SNP therein has a p-value below the genome-wide significance threshold. Even though it may be tempting to expect this method to control the family-wise error rate (and by extension, the FDR), the results from Figure 8 suggest that this is not the case, and we confirm this directly in Figure 9. Similarly, we conclude that all of the other less conservative benchmarks from Section 5.2 will fail to control the FDR in this setting. By contrast, the Local Digital Twin Test does not suffer from this problem because its p-values are valid for the conditional null hypotheses in (1). Indeed, Figure 9 shows that the Local Digital Twin Test controls the FDR at the nominal level.

6 Analysis of autism spectrum disorders

We apply the Digital Twin Test to study the genetic basis of autism spectrum disorder (ASD) using a data set of 2,565 parent-child trios from the Autism Sequencing Consortium (ASC) [60], accessed through dbGaP [61, 62]; see Appendix E for details about the sample and data processing. ASD has complex genetic roots, with variability arising from common SNPs, copy number variation, and de novo gene-disrupting mutations [63, 64]. It has been theoretically conjectured [65] and empirically observed [66] that common variants only have small effects on ASD. As a result, only recently have individual SNPs been implicated by GWAS, and the first individual SNPs appear in [67], where 5 SNPs were reported at the genome-wide significance level in a sample of approximately 45,000 individuals. Since one of these five SNPs, *rs910805* (an intergenic variant on chromosome 20), was also genotyped in the ASC trio data, we will test its causal validity with the Local Digital Twin Test. Note, however, that the ASC trio data was used as part of [67], so our results here must be viewed as a practical demonstration and not as an independent replication of these findings.

Since we do not have a corresponding cohort of non-trio observations, we use the Local Digital Twin Test with the test statistic in (6) and report the results in Table 1. We test groups centered around SNP rs910805 of increasing size, ranging from 1 Mb to the whole chromosome (in which case this Local Digital Twin Test is equivalent to the TDT). For large enough groups, we find significance at the $\alpha = 0.05$ level—we are only testing one SNP so we do not need to achieve the genome-wide significance level. We are unable to reject at finer resolutions; smaller groups correspond to fundamentally more demanding statistical hypotheses, and the effect size here is small. Our analysis suggests that the observed association with rs910805 is not due to confounding and that there is instead a causal variant in its vicinity.

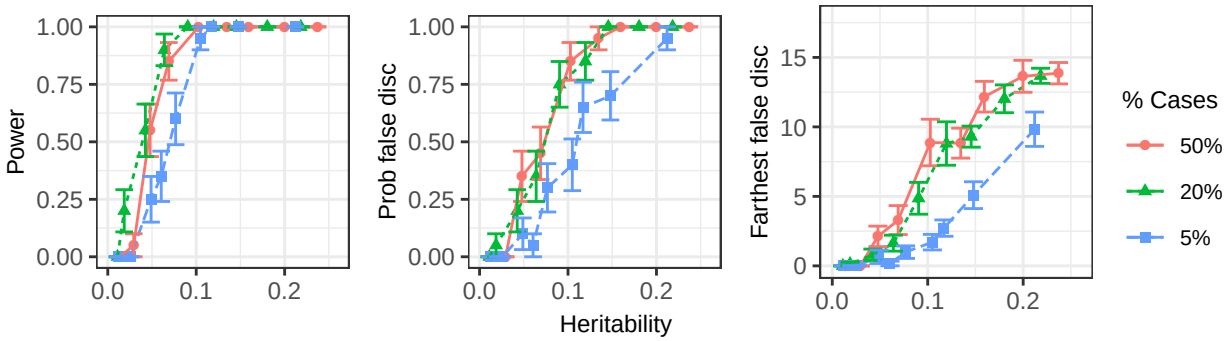

Figure 8: Performance of the TDT in an admixed population. Left: power. Middle: probability of reporting at least one spurious discovery at least 1 Mb away from the causal SNP. Right: distance from the true causal SNP to the farthest spurious discovery.

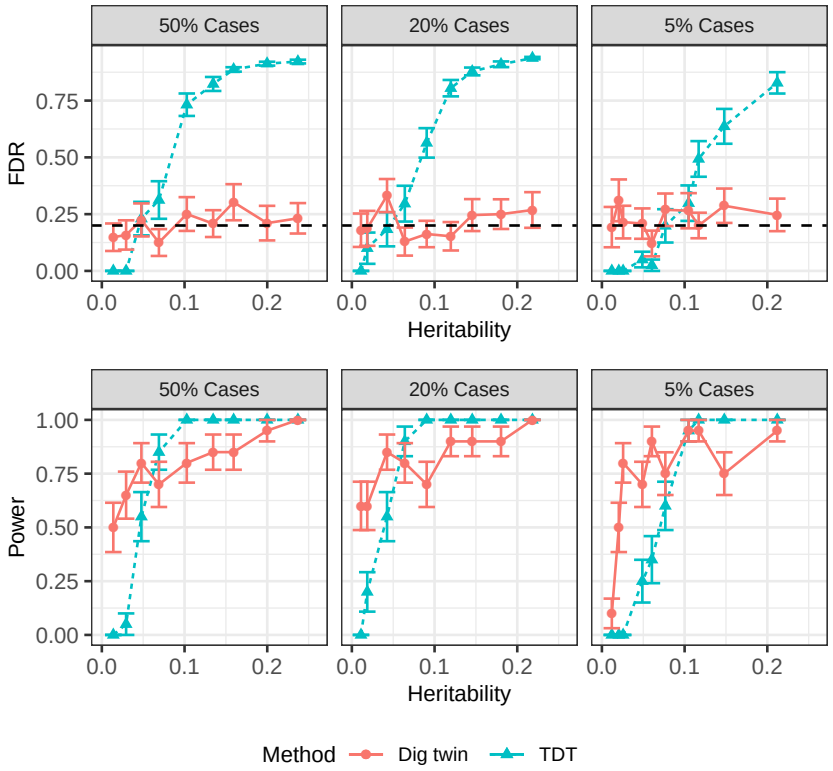

Figure 9: Performance of the Local Digital Twin Test and TDT in an admixed population. In the top panel, the dashed horizontal line indicates the nominal FDR level for the Local Digital Twin Test. Because the TDT is using the genome-wide significance level, the nominal FDR level for the TDT is less than 0.05.

Resolution	1 Mb	2 Mb	3 Mb	4 Mb	5 Mb	full-chromomsome
p-value	0.237	0.146	0.100	0.0168	0.0244	0.011

Table 1: Analysis of ASD with Local Digital Twin Test at different resolutions.

7 Discussion

We have developed a flexible statistical test to rigorously establish that a genomic region contains causal variants. It requires no modeling assumptions because the inferences are based only on variation arising in meiosis, a randomized experiment performed by nature. Nonetheless, we highlight here the two most important limitations of this work. First, our method currently relies on computationally phased haplotypes, as discussed in Section 2.1. Even though phasing is typically accurate, especially when family data are available, it would be interesting to thoroughly understand the robustness to phasing errors. Second, in order to operate at practical resolutions, our method requires many parent-child trios or duos, which are more challenging to collect than unrelated individuals. Fortunately, larger data sets will become more abundant as sequencing costs are rapidly decreasing [68]. Moreover, contemporary studies genotyping entire populations (e.g., in Iceland [69] and Finland [70]) are particularly promising because they simultaneously address the challenges of accurate phasing and obtaining data from both parents and offspring. In such studies, drawing conclusions from GWAS that are provably immune to confounding variables is appealing, making the Digital Twin Test a tantalizing way to confirm candidate GWAS discoveries.

Turning to the statistical aspect of this work, our finite-sample causal conclusions come from the fact that (i) A blocks the relevant confounders, and (ii) we exactly know the distribution of X given A. This statistical form may appear in other settings. Concretely, for any random variables X, Y and Z such that the relationship in Definition 1 holds, Theorem 1 holds as well. More generally, for a structural equation model on vectors (A, X, Y, Z), if A satisfies the back-door criterion [31] with respect to X and Y, then tests of the hypothesis (1) only detect causal effects. One can then leverage conditional testing techniques, such as the conditional randomization test or knockoffs, to make causal inferences with finite-sample guarantees.

Acknowledgements

S. B. is partly supported by a Ric Weiland Fellowship. S. B., M. S., C. S., and E. C. are supported by NSF grant DMS 1712800. C. S. and E. C. are also supported by a Math+X grant (Simons Foundation) and by NSF grant 1934578. We thank Trevor Hastie, Manuel Rivas, Robert Tibshirani, and Wing Wong for discussions of an early version of this manuscript and the Stanford Research Computing Center for computational resources and support.

References

[1] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. Mccarthy, M. A. Brown, and J. Yang, "10 years of GWAS discovery: Biology, function, and translation," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, 2017.

[2] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

[3] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nat. Genet.*, vol. 38, no. 8, pp. 904–909, 2006.

[4] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, N. B. Freimer, C. Sabatti, E. Eskin, *et al.*, "Variance component model to account for sample structure in genome-wide association studies," *Nat. Genet.*, vol. 42, pp. 348–354, 2010.

[5] D. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *J. Educ. Psychol.*, vol. 66, no. 5, pp. 688–701, 1974.

[6] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *Am. J. Hum. Genet.*, vol. 52, no. 3, pp. 506–16, 1993.

[7] R. S. Spielman and W. J. Ewens, "The TDT and other family-based tests for linkage disequilibrium and association," *Am. J. Hum. Genet.*, vol. 59, no. 5, pp. 983–989, 1996.

[8] P. Rubinstein, M. Walker, C. Carpenter, C. Carrier, J. Kressner, C. Falk, and C. F., "Genetics of HLA-disease associations. the use of the haplotype relative risk (HRR) and the 'haplo-delta' (Dh) estimates in juvenile diabetes from three racial groups," *Hum. Immunol.*, vol. 3, no. 4, p. 384, 1981.

[9] C. T. Falk and P. Rubinstein, "Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations," *Ann. Hum. Genet.*, vol. 51, no. 3, pp. 227–233, 1987.

[10] J. Ott, Y. Kamatani, and M. Lathrop, "Family-based designs for genome-wide association studies," *Nat. Rev. Genet.*, vol. 12, no. 7, pp. 465–474, 2011.

[11] G. Thomson, "Mapping disease genes: family-based association studies," *Am. J. Hum. Genet.*, vol. 57, no. 2, pp. 487–98, 1995.

[12] D. Rabinowitz and N. Laird, "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information," *Hum. Hered.*, vol. 50, no. 4, pp. 211–223, 2000.

[13] C. Lange and N. M. Laird, "Power calculations for a general class of family-based association tests: Dichotomous traits," *Am. J. Hum. Genet.*, vol. 71, no. 3, pp. 575–584, 2002.

[14] L. C. Lazzeroni and K. Lange, "A conditional inference framework for extending the transmission/disequilibrium test," *Hum. Hered.*, vol. 48, no. 2, pp. 67–81, 1998.

[15] X. Xu, C. S. Rakovski, X. Xu, and N. M. Laird, "An efficient family-based association test using multiple markers," *Genet. Epidemiol.*, vol. 31, no. 8, pp. 789–796, 2007.

[16] C. S. Rakovski, S. T. Weiss, N. M. Laird, and C. Lange, "FBAT-SNP-PC: an approach for multiple markers and single trait in family-based association tests," *Hum. Hered.*, vol. 66, no. 2, pp. 122–126, 2008.

[17] N. M. Laird and C. Lange, "The role of family-based designs in genome-wide association studies," *Stat. Sci.*, vol. 24, no. 4, pp. 388–397, 2010.

[18] D. B. Allison, "Transmission-disequilibrium tests for quantitative traits," *Am. J. Hum. Genet.*, vol. 60, no. 3, pp. 676–690, 1997.

[19] D. Rabinowitz, "A transmission disequilibrium test for quantitative trait loci," *Hum. Hered.*, vol. 47, no. 6, pp. 342–350, 1997.

[20] S. A. Monks and N. L. Kaplan, "Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus," *Am. J. Hum. Genet.*, vol. 66, no. 2, pp. 576–592, 2000.

[21] W. J. Ewens, M. Li, and R. S. Spielman, "A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker," *PLOS Genetics*, vol. 4, no. 9, pp. 1–6, 2008.

[22] D. J. Schaid, "General score tests for associations of genetic markers with disease using cases and their parents," *Genet. Epidemiol.*, vol. 13, no. 5, pp. 423–449, 1996.

[23] H. J. Cordell and D. G. Clayton, "A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes," *Am. J. Hum. Genet.*, vol. 70, no. 1, pp. 124–141, 2002.

[24] K. Van Steen, M. B. McQueen, A. Herbert, B. Raby, H. Lyon, D. L. DeMeo, A. Murphy, J. Su, S. Datta, C. Rosenow, *et al.*, "Genomic screening and replication using the same data set in family-based association testing," *Nat. Genet.*, vol. 37, no. 7, pp. 683–691, 2005.

[25] R. Fisher, *Statistical methods for research workers.* Edinburgh: Oliver & Boyd, 1925.

[26] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, "On the application of probability theory to agricultural experiments. essay on principles. section 9," *Stat. Sci.*, vol. 5, no. 4, pp. 465–472, 1990.

[27] R. Fisher, *The Design of Experiments.* Edinburgh: Oliver and Boyd, 1935.

[28] D. B. Rubin, "Causal inference using potential outcomes," *J. Am. Stat. Assoc.*, vol. 100, no. 469, pp. 322–331, 2005.

[29] K. A. Bollen, *Structural Equations with Latent Variables.* New York: Wiley, 1989.

[30] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, (USA), p. 255270, Elsevier Science Inc., 1990.

[31] J. Pearl, *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2nd ed., 2009.

[32] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search.* MIT press, 2nd ed., 2000.

[33] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *J. Mach. Learn. Res.*, vol. 8, pp. 613–636, 2007.

[34] M. H. Maathuis, M. Kalisch, and P. Bhlmann, "Estimating high-dimensional intervention effects from observational data," *Ann. Statist.*, vol. 37, pp. 3133–3164, 12 2009.

[35] D. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, 2002.

[36] Y. He and Z. Geng, "Active learning of causal networks with intervention experiments and optimal designs," *Journal of Machine Learning Research*, vol. 9, pp. 2523–2547, 11 2008.

[37] E. Candès, Y. Fan, L. Janson, and J. Lv, "Panning for gold: model-X knockoffs for high dimensional controlled variable selection," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 80, no. 3, pp. 551–577, 2018.

[38] R. F. Barber and E. Candès, "Controlling the false discovery rate via knockoffs," *Ann. Stat.*, vol. 43, no. 5, pp. 2055–2085, 2015.

[39] M. Sesia, C. Sabatti, and E. Candès, "Gene hunting with hidden markov model knockoffs," *Biometrika*, vol. 106, no. 1, pp. 1–18, 2019.

[40] M. Sesia, C. Sabatti, and E. Candès, "Rejoinder: 'Gene hunting with hidden Markov model knockoffs'," *Biometrika*, vol. 106, pp. 35–45, 02 2019.

[41] M. Sesia, E. Katsevich, S. Bates, E. Candès, and C. Sabatti, "Multi-resolution localization of causal variants across the genome," *bioRxiv preprint*, 2019. https://doi.org/10.1101/631390.

[42] M. Katan, "Apoupoprotein e isoforms, serum cholesterol, and cancer," *Lancet*, vol. 327, no. 8479, pp. 507–508, 1986.

[43] G. Davey Smith and S. Ebrahim, "Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease?," *Int. J. Epidemiol.*, vol. 32, no. 1, pp. 1–22, 2003.

[44] D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping," *Nat. Rev. Gen.*, vol. 19, no. 8, pp. 491–504, 2018.

[45] S. R. Browning and B. L. Browning, "Haplotype phasing: Existing methods and new developments," *Nat. Rev. Genet.*, vol. 12, no. 10, pp. 703–714, 2011.

[46] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, G. R. Abecasis, and P. Donnelly, "A comparison of phasing algorithms for trios and unrelated individuals," *Am. J. Hum. Genet.*, vol. 78, no. 3, pp. 437–450, 2006.

[47] J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, *et al.*, "A general approach for haplotype phasing across the full spectrum of relatedness," *PLoS Genet.*, vol. 10, no. 4, 2014.

[48] J. Haldane, "The combination of linkage values and the calculation of distances between the loci of linked factors," *J. Genet.*, vol. 8, no. 29, pp. 299–309, 1919.

[49] D. Altshuler, P. Donnelly, I. H. Consortium, *et al.*, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, p. 1299, 2005.

[50] C. Bherer, C. L. Campbell, and A. Auton, "Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales," *Nat. Comm.*, vol. 8, 2017.

[51] "New insights into the generation and role of de novo mutations in health and disease," *Geno. Bio.*, vol. 17, no. 1, p. 241, 2016.

[52] W. Tansey, V. Veitch, H. Zhang, R. Rabadan, and D. M. Blei, "The holdout randomization test: Principled and easy black box feature selection," *arXiv preprint*, 2018. arXiv:1811.00645.

[53] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini, "The UK Biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[54] X. Wang, R. Xiao, X. Zhu, and M. Li, "Gene mapping in admixed families: a cautionary note on the interpretation of the transmission disequilibrium test and a possible solution," *Hum. Hered.*, vol. 81, no. 2, pp. 106–116, 2016.

[55] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[56] R. Dai and R. Barber, "The knockoff filter for FDR control in group-sparse and multitask regression," in *International Conference on Machine Learning*, pp. 1851–1859, 2016.

[57] K. Roeder and L. Wasserman, "Genome-wide significance levels and weighted hypothesis testing," *Stat. Sci.*, vol. 24, no. 4, pp. 398–413, 2009.

[58] Z. Ren and E. Candès, "Knockoffs with side information," *arXiv preprint*, 2020. arXiv:2001.07835.

[59] A. Li and R. F. Barber, "Accumulation tests for FDR control in ordered hypothesis testing," *J. Am. Stat. Assoc.*, vol. 112, no. 518, pp. 837–849, 2017.

[60] J. D. Buxbaum, M. J. Daly, B. Devlin, T. Lehner, K. Roeder, M. W. State, A. S. Consortium, *et al.*, "The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders," *Neuron*, vol. 76, no. 6, pp. 1052–1056, 2012.

[61] M. D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nat. Genet.*, vol. 39, no. 10, pp. 1181–1186, 2007.

[62] K. A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, *et al.*, "NCBI's database of genotypes and phenotypes: dbGaP," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 975–979, 2014.

[63] I. Iossifov, B. J. Oroak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, *et al.*, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[64] T. Gaugler, L. Klei, S. J. Sanders, C. A. Bodea, A. P. Goldberg, A. B. Lee, M. Mahajan, D. Manaa, Y. Pawitan, J. Reichert, *et al.*, "Most genetic risk for autism resides with common variation," *Nat. Genet.*, vol. 46, no. 8, p. 881, 2014.

[65] B. Devlin, N. Melhem, and K. Roeder, "Do common variants play a role in risk for autism? Evidence and theoretical musings," *Brain Res.*, vol. 1380, pp. 78–84, 2011.

[66] R. Anney, L. Klei, D. Pinto, J. Almeida, E. Bacchelli, G. Baird, N. Bolshakova, S. Bölte, P. F. Bolton, T. Bourgeron, *et al.*, "Individual common variants exert weak effects on the risk for autism spectrum disorders," *Human Mol. Genet.*, vol. 21, no. 21, pp. 4781–4792, 2012.

[67] J. Grove, S. Ripke, T. D. Als, M. Mattheisen, R. K. Walters, H. Won, J. Pallesen, E. Agerbo, O. A. Andreassen, R. Anney, *et al.*, "Identification of common genetic risk variants for autism spectrum disorder," *Nat. Genet.*, vol. 51, no. 3, pp. 431–444, 2019.

[68] "DNA sequencing costs data." https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data. Accessed: 2019-11-26.

[69] "deCODE genetics." https://www.decode.com/. Accessed: 2019-12-06.

[70] "FinnGen." https://www.finngen.fi/en. Accessed: 2019-12-06.

[71] S. Bates, E. Candès, L. Janson, and W. Wang, "Metropolized Knockoff Sampling," *arXiv preprint*, 2019. arXiv:1903.00434.

[72] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. USA: Prentice Hall PTR, 1st ed., 2000.

[73] D. B. Rubin, "Comment on "Randomization analysis of experimental data: The Fisher randomization test"," *J. Am. Stat. Assoc.*, vol. 75, no. 371, pp. 591–593, 1980.

[74] O. Delaneau, J. Marchini, and J.-F. Zagury, "A linear complexity phasing method for thousands of genomes," *Nat. Methods*, vol. 9, no. 2, pp. 179–181, 2012.

[75] A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[76] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz, *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015.

[77] D. S. Falconer, "The inheritance of liability to certain diseases, estimated from the incidence among relatives," *Ann. Hum. Gen.*, vol. 29, no. 1, pp. 51–76, 1965.

[78] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: rising to the challenge of larger and richer datasets," *GigaScience*, vol. 4, no. 1, 2015.

[79] D. Fulker, S. Cherny, P. Sham, and J. Hewitt, "Combined linkage and association sib-pair analysis for quantitative traits," *Am. J. Hum. Genet.*, vol. 64, no. 1, pp. 259–267, 1999.

[80] G. R. Abecasis, L. R. Cardon, and W. Cookson, "A general test of association for quantitative traits in nuclear families," *Am. J. Hum. Genet.*, vol. 66, no. 1, pp. 279–292, 2000.

[81] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, pp. 559–575, 2007.

A Constructing independent p-values

We present a technical refinement of the Local Digital Twin Test (still testing the null hypothesis in (8)) that produces independent p-values for disjoint null groups. The core idea is that for each individual, we determine which groups contain a recombination event; then, we generate the digital twins only in those groups, sampling from a modified version of the distribution (9). By construction, these digital twins will have recombination events in the same groups as the true observations, but the recombination events will be randomly perturbed. This separates the randomness used to test each region. We then show how to create feature importance statistics that ignore the randomness in other regions. As a result, the tests in different regions are fully decoupled.

To begin, let G denote the collection of disjoint groups of SNPs that we wish to test. We will assume that each group g is a continuous block of the form $\{g_-, g_- + 1, \ldots, g_+\}$ for endpoints $g_- \leq g_+$ in $\{1, \ldots, p\}$. Also, let $\mathcal{B} \subset \{1, \ldots, p\}$ be the set of SNPs that form the boundary of some group: $\mathcal{B} = \{j : j = g_- \text{ or } j = g_+ \text{ for some } g \in G\}$.

For simplicity, we consider a single observation and define $U^m \in \{a, b\}^p$ and $U^f \in \{a, b\}^p$ to be the underlying ancestral states for the two haplotypes X^f and X^m. We first sample from the posterior distribution conditional on the data based on the HMM described in Section 2.5 and then condition on $U^m_{\mathcal{B}}$ and $U^f_{\mathcal{B}}$. This conditioning splits the distribution of (X^f_g, U^f_g) and (X^m_g, U^m_g) into *independent* HMMs across the groups $g \in G$, see [71] for a general discussion of this conditioning idea. In the modified Local Digital Twin Test, each digital twin will be sampled from the distribution of

$$(X^f_g, X^m_g) \mid (U^m_{\mathcal{B}}, U^f_{\mathcal{B}}, A). \tag{13}$$

This is essentially the same as in the original Local Digital Twin Test, which instead samples from (9), because the values $U^m_{\mathcal{B}}, U^f_{\mathcal{B}}$ are almost perfectly determined by the data X^f_{-g} and X^m_{-g}, since the measured SNPs are dense and recombination is rare. Sampling from the distribution (13) is straightforward: we simply sample U^m_g and U^f_g, an HMM sampling operation, and then from this we sample X^m_g and X^f_g from the emission distribution.

We now introduce additional notation that will be needed to describe the digital twin sampling. Given $U^{f,(\mathrm{i})}$ and $U^{m,(\mathrm{i})}$, for each observation $i = 1, \ldots, n$, let $R^{m,(\mathrm{i})} = \{g \in G : U^{m,(\mathrm{i})}_{g_-} \neq U^{m,(\mathrm{i})}_{g_+}\}$ be the set of groups where individual i has an odd number of recombinations (for small groups, there will typically be exactly one recombination). Define $R^{f,(\mathrm{i})}$ in the analogous way. Now, let $\mathcal{D}^m = \{(i, j) : j \notin g \text{ for all } g \in R^{m,(\mathrm{i})}\}$ be the set of observations and sites in groups with an even number of recombinations. Define $\mathcal{D}^f$ in the analogous way. The upcoming algorithm will hold $X^{m,(\mathrm{i})}_j$ for $(i, j) \in \mathcal{D}^m$ fixed when generating the digital twins.

Next, we define the masked version of X which is safe to use in the feature importance statistics; feature importance statistics using the masked version of X will operate independently in different groups. The masked version is defined as follows:

$$\begin{aligned} X^{(\mathrm{mask,m})(\mathrm{i})}_j &= \begin{cases} X^{\mathrm{m},(\mathrm{i})}_g & \text{if } (i, j) \in \mathcal{D}^m, \\ E[X^{\mathrm{m},(\mathrm{i})}_g \mid M_a, M_b, U^m_{\mathcal{B}}] & \text{otherwise,} \end{cases} \\ X^{(\mathrm{mask,f})(\mathrm{i})}_j &= \begin{cases} X^{\mathrm{f},(\mathrm{i})}_g & \text{if } (i, j) \in \mathcal{D}^f, \\ E[X^{\mathrm{f},(\mathrm{i})}_g \mid M_a, M_b, U^f_{\mathcal{B}}] & \text{otherwise,} \end{cases} \end{aligned} \tag{14}$$

for each observation $i = 1, \ldots, n$. The entries of X^m in groups with no recombination events will

remain unchanged in both the digital twins and in $X^{(\text{mask,m})(\text{i})}$. On the other hand, for groups with observed recombinations, the digital twins will vary, so, for these entries, we set $X^{(\text{mask,m})}$ to be a constant: the average imputation based on the parental haplotypes and $(U^m_{\mathcal{B}}, U^f_{\mathcal{B}})$. These conditional expectations can be computed easily, thanks to the Markov property. Notice that, since there are very few groups with recombination events, most entries of $X^{(\text{mask,m})}$ will be equal to those of X^m, and most entries of $X^{(\text{mask,f})}$ will be equal to those of X^f. For instance, with groups of size 1 Mb we will have that over 99% of the entries match.

With the notation in place, we can finally present the procedure in Algorithm 3.

Algorithm 3: The Local Digital Twin Test with independent p-values

— Sample U^m and U^f given (X^m, X^f, A) according the HMM in Section 2.5 (see Appendix C for an explicit sampler).
— Define $X^{(\text{mask,m})}$ and $X^{(\text{mask,f})}$ according to (14).
for $g \in G$ **do**
  — Compute the test statistic on the true data:

$$t^* = T(X^{(\text{mask,m})}_{-g} + X^{(\text{mask,f})}_{-g}, X^m_g + X^f_g, Y).$$

  for $k = 1, \ldots, K$ **do**
    — For observations i such that $g \in R^{\text{m,(i)}}$, sample $\widetilde{X}^{\text{m,(i)}}_g$ from the distribution (13), independent of X^f, X^m, U^f, U^m and Y (see Appendix C for an explicit sampler). Otherwise, set $\widetilde{X}^{\text{m,(i)}}_g = X^{\text{m,(i)}}_g$. Sample $\widetilde{X}^f_g$ analogously.
    — Compute the test statistic on the digital twins:

$$t_k = T(X^{(\text{mask,m})}_{-g} + X^{(\text{mask,f})}_{-g}, \widetilde{X}^m_g + \widetilde{X}^f_g, Y).$$

  end
  — Compute the quantile of the true statistic t^* among the digital twin statistics $t_1, \ldots, t_K$:

$$v_g = \frac{1 + \#\{k : t^* \leq t_k\}}{K + 1}.$$

end
Result: A family of independent p-values $\{v_g\}_{g \in G}$ for testing the conditional hypotheses (8).

The null p-values produced by this algorithm are jointly independent, which we record in Theorem 2 in Section 4.2. The proof is given in Appendix B.

Remark This procedure can be viewed as an analogue to the model-X knockoffs procedure [37], except with K knockoffs instead of the usual two. Conditional on $U_{\mathcal{B}}$, the Markov structure ensures that the imputations (i.e. the multiple knockoffs) are jointly exchangeable and hence are analogous to multiple knockoffs. Masking the values of X where the imputations take place is analogous to making the feature importance statistic for group g invariant to swapping $X_{g'}$ and $\widetilde{X}_{g'}$ for all groups g' in the original knockoff setting.

B Proofs

We first give an explicit proof for Proposition 1. The argument is essentially the same as that used in Section 2.2 to prove Theorem 1.

Proof of Proposition 1. Suppose the null hypothesis H_0 is true, so that

$$Y \perp\!\!\!\perp X_{\mathcal{C}} \mid (A, Z).$$

Then, since Z is an external confounder,

$$Z \perp\!\!\!\perp X_{\mathcal{C}} \mid A,$$

and so

$$(Y, Z) \perp\!\!\!\perp X_{\mathcal{C}} \mid A.$$

Thus, the following null hypothesis is also true:

$$H_0' : Y \perp\!\!\!\perp X_{\mathcal{C}} \mid A.$$

The Digital Twin Test is a valid test of H_0', since it is an instance of the conditional randomization test [37], and so the result follows. □

Next, we turn to Proposition 2. The proof is essentially the same as the above.

Proof of Proposition 2. Suppose the null hypothesis H_0^g is true so that

$$Y \perp\!\!\!\perp X_g \mid (X_{-g}^m, X_{-g}^f, A, Z).$$

Then, since Z is an external confounder, we have that

$$Z \perp\!\!\!\perp X_g \mid (X_{-g}^m, X_{-g}^f, A),$$

and so

$$(Y, Z) \perp\!\!\!\perp X_g \mid (X_{-g}^m, X_{-g}^f, A).$$

Thus, the following null hypothesis is also true:

$$H_0'^g : Y \perp\!\!\!\perp X_g \mid (X_{-g}^m, X_{-g}^f, A).$$

The Local Digital Twin Test is a valid test of $H_0'^g$, since it is an instance of the conditional randomization test [37], and so the result follows. □

Lastly, we give a proof that the p-values from the modified Local Digital Twin Test in Appendix A are super-uniform and independent.

Proof of Theorem 2. First, note that any group g satisfying the Local Digital Twin Test null hypothesis in (10) satisfies the weaker null hypothesis in (8) by the argument in the proof of Proposition 2. As a result, it suffices to show the result for the case without any external confounder Z.

Let $n_G = |G|$ be the number of groups and recall that K is the number of iterations in Algorithm 3. Conditional on $(U^m_{\mathcal{B}}, U^f_{\mathcal{B}}, A)$, we have the following simultaneous exchangeability property:

$$\Big((X^m_{g_1}, \widetilde{X}^{m,1}_{g_1}, \ldots, \widetilde{X}^{m,K}_{g_1}), \ldots, (X^m_{g_{n_G}}, \widetilde{X}^{m,1}_{g_{n_G}}, \ldots, \widetilde{X}^{m,K}_{g_{n_G}})\Big) \overset{d}{=} \\ \Big((X^m_{g_1}, \widetilde{X}^{m,1}_{g_1}, \ldots, \widetilde{X}^{m,K}_{n_G})_{\sigma_1}, \ldots, (X^m_{g_{n_G}}, \widetilde{X}^{m,1}_{g_{n_G}}, \ldots, \widetilde{X}^{m,K}_{g_{n_G}})_{\sigma_{n_G}}\Big), \quad (15)$$

for any permutations $\sigma_1, \ldots, \sigma_{n_G}$ of $(1, \ldots, K+1)$. This is true because conditional on $(U^m_{\mathcal{B}}, U^f_{\mathcal{B}}, A)$, the X^m_g for different groups g are independent HMMs. Moreover, (15) continues to hold if we additionally condition on $(X^{(\text{mask,m})}, X^{(\text{mask,f})})$, since this is equivalent to conditioning on more entries of U^f and U^m, after which the X^m_g remain independent HMMs. The analogous property holds for X^f.

Next, we show that (15) continues to hold for the null groups even after additionally conditioning on Y. Without loss of generality, suppose g_1 is a null group. First, by construction of (U^m, U^f) we have that

$$Y \perp\!\!\!\perp (U^m, U^f) \mid X^f, X^m, A.$$

As a result,

$$Y \perp\!\!\!\perp X_{g_1} \mid (X^m_{-g_1}, X^f_{-g_1}, A) \implies Y \perp\!\!\!\perp (X_{g_1}, U^m, U^f) \mid (X^m_{-g_1}, X^f_{-g_1}, A).$$

The right-hand side of the above further implies that

$$Y \perp\!\!\!\perp X_{g_1} \mid (X^m_{-g_1}, X^f_{-g_1}, X^{(\text{mask,m})}, X^{(\text{mask,f})}, U^f_{\mathcal{B}}, U^m_{\mathcal{B}}, A),$$

since $(X^{(\text{mask,m})}, X^{(\text{mask,f})}, U^f_{\mathcal{B}}, U^m_{\mathcal{B}})$ is a function of $(X^m_{g_1}, X^f_{g_1}, X^m_{-g_1}, X^f_{-g_1}, U^f, U^m, A)$. By the independence in the construction of the digital twins, we then have

$$Y \perp\!\!\!\perp X_{g_1} \mid \big((X^m_{g_2}, \widetilde{X}^{m,1}_{g_2}, \ldots, \widetilde{X}^{m,K}_{g_2}), \ldots, (X^m_{g_{n_G}}, \widetilde{X}^{m,1}_{g_{n_G}}, \ldots, \widetilde{X}^{m,K}_{g_{n_G}}), \\ X^{(\text{mask,m})}, X^{(\text{mask,f})}, U^f_{\mathcal{B}}, U^m_{\mathcal{B}}, A\big),$$

and further

$$Y \perp\!\!\!\perp (X^m_{g_1}, \widetilde{X}^{m,1}_{g_1}, \ldots, \widetilde{X}^{m,K}_{g_1}) \mid \Big((X^m_{g_2}, \widetilde{X}^{m,1}_{g_2}, \ldots, \widetilde{X}^{m,K}_{g_2}), \ldots, (X^m_{g_{n_G}}, \widetilde{X}^{m,1}_{g_{n_G}}, \ldots, \widetilde{X}^{m,K}_{g_{n_G}}), \\ X^{(\text{mask,m})}, X^{(\text{mask,f})}, U^f_{\mathcal{B}}, U^m_{\mathcal{B}}, A\Big).$$

We conclude that (15) holds conditional on $(X^{(\text{mask,m})}, X^{(\text{mask,f})}, U^f_{\mathcal{B}}, U^m_{\mathcal{B}}, A, Y)$ for any permutations $\sigma_1, \ldots, \sigma_{n_G}$ of $(1, \ldots, K+1)$ such that the non-null groups have the identity permutation. The analogous statement holds for X^f. Combining this with the analogous result for X^f, we then have:

$$\Big((X_{g_1}, \widetilde{X}^1_{g_1}, \ldots, \widetilde{X}^K_{g_1}), \ldots, (X_{g_{n_G}}, \widetilde{X}^1_{g_{n_G}}, \ldots, \widetilde{X}^K_{g_{n_G}})\Big) \overset{d}{=} \\ \Big((X_{g_1}, \widetilde{X}^1_{g_1}, \ldots, \widetilde{X}^K_{n_G})_{\sigma_1}, \ldots, (X_{g_{n_G}}, \widetilde{X}^1_{g_{n_G}}, \ldots, \widetilde{X}^K_{g_{n_G}})_{\sigma_{n_G}}\Big), \quad (16)$$

conditional on $(X^{(\text{mask,m})}, X^{(\text{mask,f})}, U^f_{\mathcal{B}}, U^m_{\mathcal{B}}, A, Y)$ for any permutations $\sigma_1, \ldots, \sigma_{n_G}$ of $(1, \ldots, K+1)$ such that the non-null groups have the identity permutation.

We now turn to the feature importance statistics $t^*, t_1, \ldots, t_K$ for group g. These are a function of $(X_g^m, \widetilde{X}_g^{m,1}, \ldots, \widetilde{X}_g^{m,K})$ and $(X_g^f, \widetilde{X}_g^{f,1}, \ldots, \widetilde{X}_g^{f,K})$ together with $(X^{(\text{mask,m})}, X^{(\text{mask,f})}, U_{\mathcal{B}}^f, U_{\mathcal{B}}^m, A, Y)$, and we are conditioning on the latter. In particular, the feature importance statistics for group g do not depend on $(X_{g'}^m, \widetilde{X}_{g'}^{m,1}, \ldots, \widetilde{X}_{g'}^{m,k})$ for any other group g', except through the entries already fixed in $X^{(\text{mask,m})}$. Combining this fact with the result (16), we have that

$$\left((t^*_{g_1}, t^1_{g_1}, \ldots, t^K_{g_1}), \ldots, (t^*_{g_{n_G}}, t^1_{g_{n_G}}, \ldots, t^K_{g_{n_G}})\right) \stackrel{d}{=} \left((t^*_{g_1}, t^1_{g_1}, \ldots, t^K_{g_1})_{\sigma_1}, \ldots, (t^*_{g_{n_G}}, t^1_{g_{n_G}}, \ldots, t^K_{g_{n_G}})_{\sigma_{n_G}}\right),$$

for any permutations $\sigma_1, \ldots, \sigma_{n_G}$ of $(1, \ldots, K+1)$, provided the non-null groups have the identity permutation, conditional on $(X^{(\text{mask,m})}, X^{(\text{mask,f})}, U_{\mathcal{B}}^f, U_{\mathcal{B}}^m, A, Y)$. This implies that conditionally the p-values v_g for all null groups g are jointly independent and have the discrete uniform distribution (if we randomly break ties), and so the same holds after marginalizing out $(X^{(\text{mask,m})}, X^{(\text{mask,f})}, U_{\mathcal{B}}^f, U_{\mathcal{B}}^m, Y)$. □

C Sampling digital twins

In this section, we explicitly describe how to sample the digital twins in Algorithms 1-3. In each case, the sampling is a standard conditional sampling operation on an HMM [e.g., 72], but we give the details for our setting for completeness.

For convenience, we consider a single observation so that $X^m, X^f, M^a, M^b, F^a, F^b \in \{0,1\}^p$ are the haplotypes, and whose hidden ancestral states are $U^m, U^f \in \{a,b\}^p$. Without loss of generality, we consider one chromosome; different chromosomes are independent, so this is sufficient. In addition, for $u \in \{a,b\}$ we define:

$$\overline{u} = \begin{cases} a & \text{when } u = b, \\ b & \text{when } u = a, \end{cases}$$

i.e., the complementary ancestral strand.

C.1 Global digital twins

This subsection describes how to sample the digital twins in Algorithm 1. Without loss of generality, we consider X^m, sampled as follows:

1. (Hidden ancestral states) Sample u_1^m uniformly from $\{a,b\}$.
2. For $j = 1, \ldots, p$, independently sample u_j^m as

$$u_j^m = \begin{cases} u_{j-1}^m & \text{with probability } 1/2(1+e^{-2d_j}), \\ \overline{u_{j-1}^m} & \text{otherwise.} \end{cases}$$

3. (De novo mutations) For each $j = 1, \ldots, p$, independently sample X_j^m as

$$X_j^m = \begin{cases} M_j^{u_j^m} & \text{with probability } 1-\epsilon, \\ 1 - M_j^{u_j^m} & \text{otherwise.} \end{cases}$$

C.2 Local digital twins

This subsection shows how to sample the digital twins in Algorithm 2. Again, we consider only X^m explicitly.

Computing forward-backward weights

We start by computing the forward-backward weights for the HMM from Section 2.5. Recall that our model gives the following quantities:

$$P(U_j^m = u_j^m \mid U_{j-1}^m = u_{j-1}^m) = \begin{cases} \frac{1}{2}(1 + e^{-2d_j}) & \text{if } u_j^m = u_{j-1}^m, \\ 1 - \frac{1}{2}(1 + e^{-2d_j}) & \text{otherwise,} \end{cases}$$

$$P(X_j^m = x_j^m \mid U_j^m = u_j^m) = \begin{cases} 1 - \epsilon & \text{if } M_1^{u_j^m} = X_j^m, \\ \epsilon & \text{otherwise,} \end{cases}$$

and

$$P(U_j^m = U_{j'}^m) = \frac{1}{2}(1 + e^{-2(d_j + \cdots + d_{j'-1})}), \quad \text{for } j' > j.$$

We define the *forward weights* for $j = 1, \ldots, p$ as

$$\alpha_j^m(u_j^m) = P(X_{1:j}^m = x_{1:j}^m, U_j^m = u_j^m),$$

where the notation $x_{j:j'}$ indicates the subvector of x of entries $j, j+1, \ldots, j'$. The $\alpha_j^m(\cdot)$ can be computed recursively as follows:

$$\alpha_1^m(u_1^m) = \begin{cases} \frac{1}{2}(1 - \epsilon) & \text{if } M_1^{u_1^m} = X_1^m, \\ \frac{1}{2}\epsilon & \text{otherwise,} \end{cases}$$

$$\alpha_j^m(u_j^m) = \sum_{u \in \{a,b\}} \alpha_{j-1}^m(u) P(U_j^m = u_j^m \mid U_{j-1}^m = u) P(X_j^m = x_j^m \mid U_j^m = u_j^m),$$

for $j = 2, \ldots, p$ and $u_j^m \in \{a, b\}$. For convenience, we also define $\alpha_0(u) = 1/2$ for $u \in \{a, b\}$.

Next, we define the *backward weights* as

$$\beta_j^m(u_j^m) = P(X_{(j+1):p} = x_{(j+1):p} \mid U_j^m = u_j^m),$$

with the convention that $\beta_p^m(u) = 1$ for $u \in \{a, b\}$. These can be computed recursively as follows:

$$\beta_j(u_j^m) = \sum_{u \in \{a,b\}} \beta_{j+1}^m(u) P(U_{j+1}^m = u \mid U_j^m = u_j^m) P(X_{j+1}^m = x_{j+1}^m \mid U_{j+1}^m = u),$$

for $j = p - 1, \ldots, 1$ and $u_m^j \in \{a, b\}$.

Sampling the local digital twins

Consider a group of SNPs $g = \{g_-, g_- + 1, \ldots, g_+\} \subset \{1, \ldots, p\}$. We sample a local digital twin as follows:

1. (Hidden ancestral states) Sample $U^m_{g_-} \in \{a, b\}$ from the distribution

$$P(U^m_{g_-} = u) \propto P(X^m_{g_-} = x^m_{g_-} \mid U^m_{g_-} = u) \left[P(U^m_{g_+} = U^m_{g_-})\beta_{g_+}(u) + P(U^m_{g_+} \neq U^m_{g_-})\beta_{g_+}(\overline{u})\right] \\ \left[\alpha_{g_- - 1}(u)P(U^m_{g_-} = U^m_{g_- - 1}) + \alpha_{g_- - 1}(\overline{u})P(U^m_{g_-} \neq U^m_{g_- - 1})\right].$$

2. For $j = g_- + 1, \ldots, g_+$, sample $U^m_j \in \{a, b\}$ from the distribution

$$P(U^m_j = u) \propto P(X^m_j = x^m_j \mid U^m_j = u)P(U^m_j = u \mid U^m_{j-1} = u^m_{j-1}) \\ \left[P(U_{g_+} = U_j)\beta_{g_+}(u) + P(U_{g_+} \neq U_j)\beta_{g_+}(\overline{u})\right].$$

3. (De novo mutations) For each $j = g_-, \ldots, g_+$, independently sample $\widetilde{X}^m_j$ as

$$\widetilde{X}^m_j = \begin{cases} M_j^{u^m_j} & \text{with probability } 1 - \epsilon, \\ 1 - M_j^{u^m_j} & \text{otherwise.} \end{cases}$$

C.3 Modified local digital twins

This subsection describes how to sample the digital twins in Algorithm 3. As usual, we consider only X^m explicitly. First, we sample the hidden ancestral states $U^m \in \{a, b\}^p$ using the sampler from Appendix C.2, except with group $g = \{1, 2, \ldots, p\}$.

Next, we focus on a group $g = \{g_-, g_- + 1, \ldots, g_+\} \subset \{1, \ldots, p\}$. If $U^m_{g_-} = U^m_{g_+}$, then we set $\widetilde{X}^m_g = X^m_g$; otherwise, we sample the digital twin as follows:

1. Set $\widetilde{U}^m_{g_-} = U^m_{g_-}$ and $\widetilde{U}^m_{g_+} = U^m_{g_+}$.
2. Repeat until m is odd: sample $m \sim \text{Pois}(d_{g_- + 1} + \cdots + d_{g_+})$ (i.e. sample from a Poisson conditional on the outcome being odd).
3. Sample the recombination points $r_1, \ldots, r_m$ as independently from the discrete distribution supported on $\{g_- + 1, \ldots, g_+\}$ with probability vector proportional to

$$\left(\frac{1}{2}(1 + e^{-2d_{g_- + 1}}), \ldots, \frac{1}{2}(1 + e^{-2d_{g_+}})\right).$$

4. For $j = g_- + 1, \ldots, g_+$, define u^m_j as

$$u^m_j = \begin{cases} u^m_{j-1} & \text{if } \#\{k\text{: } r_k = j\} \text{ is even,} \\ \overline{u^m_{j-1}} & \text{otherwise.} \end{cases}$$

5. (De novo mutations) For each $j = g_-, \ldots, g_+ - 1$, independently sample $\widetilde{X}^m_j$ as

$$\widetilde{X}^m_j = \begin{cases} M_j^{\tilde{u}^m_j} & \text{with probability } 1 - \epsilon, \\ 1 - M_j^{\tilde{u}^m_j} & \text{otherwise.} \end{cases}$$

D Potential outcomes formulation

We can also formulate the claims of Section 2.2 in terms of the potential outcomes framework [5, 28]. Suppose that all SNPs with a causal effect are measured. This is not realistic in practical examples, but it is challenging to formulate the localization of unmeasured SNPs in the language of potential outcomes. In any case, this will be enough to demonstrate that our inferences are consistent with this point of view. Next, we assume that the potential outcomes for unit i are only affected by the subject's genotypes $x^{(\mathrm{i})}$ (the SUTVA assumption [73]) and let $Y_i(x^{(\mathrm{i})})$ denote the potential outcomes for unit i under the "treatment assignment" $x^{(\mathrm{i})} \in \{0,1,2\}^p$, and let $Y_i = Y_i(X^{(\mathrm{i})})$ denote the observed value of Y_i. Within this framework, the potential outcomes $Y_i(x^{(\mathrm{i})})$ are considered fixed (nonrandom) but unknown, whereas $X^{(\mathrm{i})}$ and Y_i are considered random. We wish to understand whether certain coordinates of our vector-valued treatment assignment have a causal effect. To formalize this, consider a group of SNPs g. We will test the sharp null hypothesis

$$H_0^{g,\mathrm{po}} : Y_i(x^{(\mathrm{i})}) = Y_i(\tilde{x}^{(\mathrm{i})}) \text{ for all } x^{(\mathrm{i})} \text{ and } \tilde{x}^{(\mathrm{i})} \text{ such that } x^{(\mathrm{i})}_{-g} = \tilde{x}^{(\mathrm{i})}_{-g}, \text{ for every unit } i.$$

That is, the potential outcomes are the same whenever the SNPs outside the group g are the same. Lastly, we assume that the treatment assignment is random, that is, that the X_i are independent samples from the recombination model in Section 2.5. Under these assumptions

$$H_0^{g,\mathrm{po}} \implies Y \perp\!\!\!\perp X_g \mid (X^m_{-g}, X^f_{-g}, A).$$

Notice that the right hand side is exactly the same as in (8), the null hypothesis tested by the Local Digital Twin Test. Thus, the Local Digital Twin Test is a valid test of $H_0^{g,\mathrm{po}}$, and the resulting inferences are causal inferences in the potential outcomes sense.

E ASD data analysis details

The ASC data was accessed through the dbGaP platform through accession phs000298.v4.p3. It contains data for 7880 individuals from 2611 pedigrees regarding $947,233$ autosomal markers. After removing 11 individuals that have high missingness and keeping only complete trios, we are left with $2,565$ trios. These genotypes are computationally phased using the `SHAPEIT` software [74] incorporating the known pedigree information [47] and using the genetic distances from the 1000 genomes project [75, 76].

F Permutation tests are not robust to confounders

Permutation tests are not a valid alternative to the Digital Twin Test for testing causal null hypotheses, as demonstrated by the following simple example. Consider a population consisting of two distinct subpopulations such that, for some given SNP X_j, $X_j = 2$ in the first subpopulation and $X_j = 0$ in the second subpopulation. Suppose we are interested in a response Y on which X_j has no causal effect whatsoever, but that happens to satisfy $Y = 1$ in the first subpopulation and $Y = 0$ in the second subpopulation due to other genetic components or external confounders. A permutation test would have no choice but to discover a significant association between Y and X_j, even though there is no causal relation. By contrast, the Digital Twin Test would never reject the causal null hypothesis because the digital twins preserve the population structure and hence satisfy $\widetilde{X}_j = X_j$, as shown in Table 2.

Indeed, it is widely known that permutation tests need to be corrected for population structure—a common approach would involve regressing the top principal components of the genetic matrix out of the response vector [3]. Although this correction may significantly mitigate the problem, it is approximate and it may not account for all other unknown confounders. By contrast, Theorem 1 tells us that the Digital Twin Test implicitly accounts for a very broad family of unknown confounders.

Subject	Population	Y	X_j	X_j^*	$\widetilde{X}_j$
1	1	1	1	0	1
2	1	1	1	1	1
3	1	1	1	1	1
4	1	1	1	0	1
5	2	0	0	1	0
6	2	0	0	0	0
7	2	0	0	1	0
8	2	0	0	0	0

Table 2: A hypothetical sample in which strong population structure creates a spurious association between the trait, Y, and some SNP, X_j. The population structure is not preserved by permutations (X_j^*), which report non-causal associations, but is preserved by the Digital Twin Test ($\widetilde{X}_j$), which can only discover causal effects.

G Auxiliary simulations and experiment details

G.1 Heritability

Throughout this paper, the results of the simulations with binary traits are indexed by the heritability, which is measured on the liability scale:

$$\frac{\mathrm{var}(\hat{Y})}{\mathrm{var}(Y)} \cdot \frac{\bar{Y}(1-\bar{Y})}{\phi(\Phi^{-1}(\bar{Y}))}.$$

Above, $\hat{Y}$ is the mean of Y conditional on the genotypes X, according to the true model, $\bar{Y}$ is the true expectation of Y, ϕ is the density of the standard normal distribution and Φ is its cumulative distribution function. This scale is widely used because it makes the range of heritability values independent of the diseases prevalence, $\bar{Y}$ [77].

G.2 Main text experiment details

For the experiment in Section 5.1, we report metrics about the lasso fit $\hat{\beta}$ on the external GWAS in Figure S.1.

Turning the experiment in Section 5.2, the hypotheses for the accumulation test are sorted in decreasing order with respect to $\sum_{j \in G_k} |\hat{\beta}_j^{\mathrm{GWAS}}(\lambda_{\mathrm{cv}})|$. We use an accumulation test with the hinge exponential function with $c = 2$ [59].

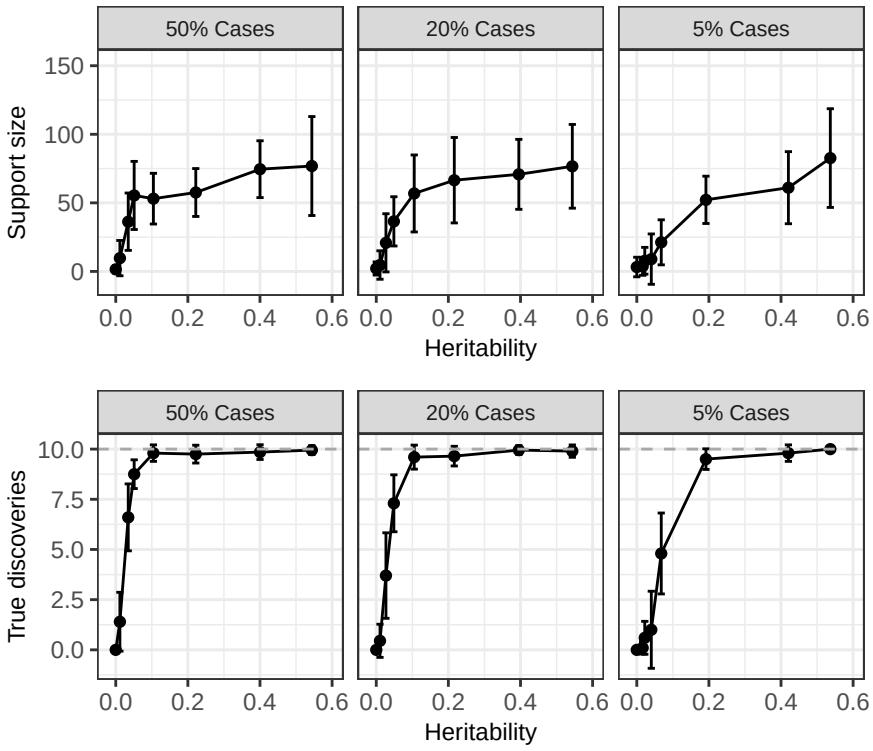

Figure S.1: Properties of the estimated regression coefficient $\hat{\beta}$ in the simulations in Section 5.1. Error bars show one standard deviation.

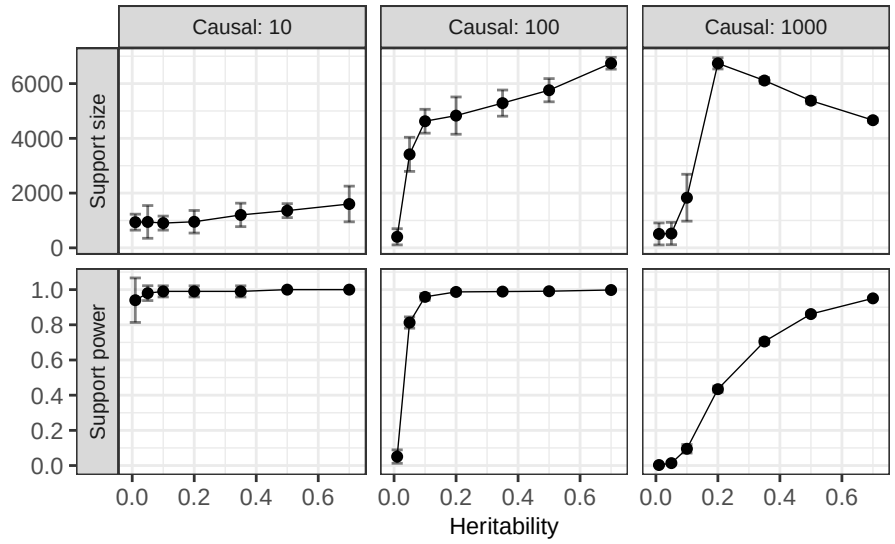

Figure S.2: Properties of the regression coefficients $\hat{\beta}$ in the full-genome simulations with binary trait. Error bars show one standard deviation.

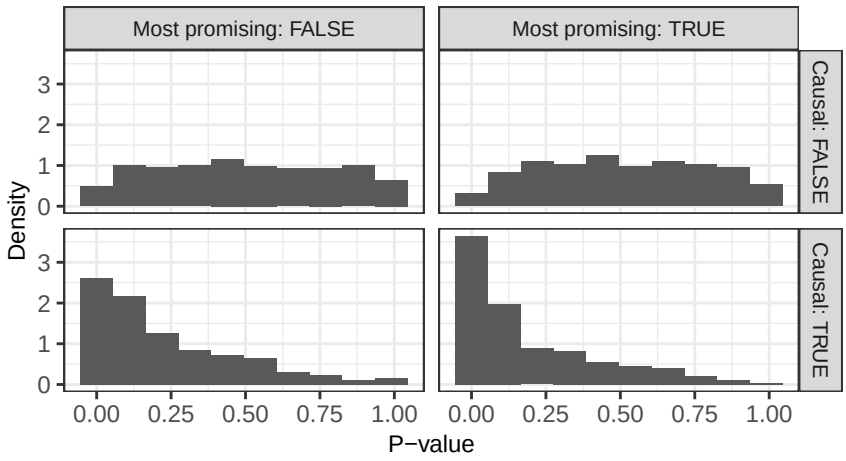

Figure S.3: Histogram of null and non-null p-values computed by the local Digital Twin Test for a simulated binary trait with 100 causal variables. The null p-values are stratified based on whether they correspond to genetic groups among the 200 most promising ones as inferred from the analysis of independent GWAS data. The total heritability of the simulated trait is 70%.

As an added diagnostic, we plot the p-values from the Local Digital Twin Test and confirm that they are uniformly distributed in Figure S.3. We confirm that the null p-values are super-uniform in distribution, and further that this holds even for the nulls that are identified as most promising by the external GWAS.[4]

Lastly, for the experiment in Section 5.3, we use the Selective SeqStep procedure [38] to combine the group p-values into a final set of discoveries. This procedure has one parameter, the threshold below which we reject p-values, which we set at $c = 0.5$.

G.3 Robustness to computational phasing

In practice, the Local Digital Twin Test relies on computationally phased haplotypes, and we now verify that it is robust to imperfect phasing. We carry out a simulation in the setting of Section 5.1, except that we split the chromosome into 30 groups of size approximately 2 Mb, and then apply the Local Digital Twin Test exactly as in Section 5.2. Note that in our synthetic population, we perfectly know the phase of both the parents and offspring. To check our robustness to phasing, we consider either using (i) the true haplotypes or (ii) computationally phased haplotypes. For the latter, we input the trio *genotypes* to the `SHAPEIT` software [74] (which accounts for the pedigree structure [47]) to obtain the computationally phased haplotypes. Lastly, we combine the group p-values using either the accumulation test [59], as in Section 5.2, or using the Selective SeqStep procedure [38] with parameter $c = 0.5$.

The results are reported in Figure S.4. We find that the power and FDR when using the computationally phased haplotypes are essentially identical to those when using the true haplotypes: the FDR is controlled at the nominal level while retaining the same power.

[4] The parameter ϵ should be set to the value 10^{-8}, the mutation rate for humans [51]. We observed that when we simulated offspring with $\epsilon = 0$ and then used a larger value (such as $\epsilon = 10^{-6}$) in the digital twin sampling, the null p-values had a non-uniform distribution.

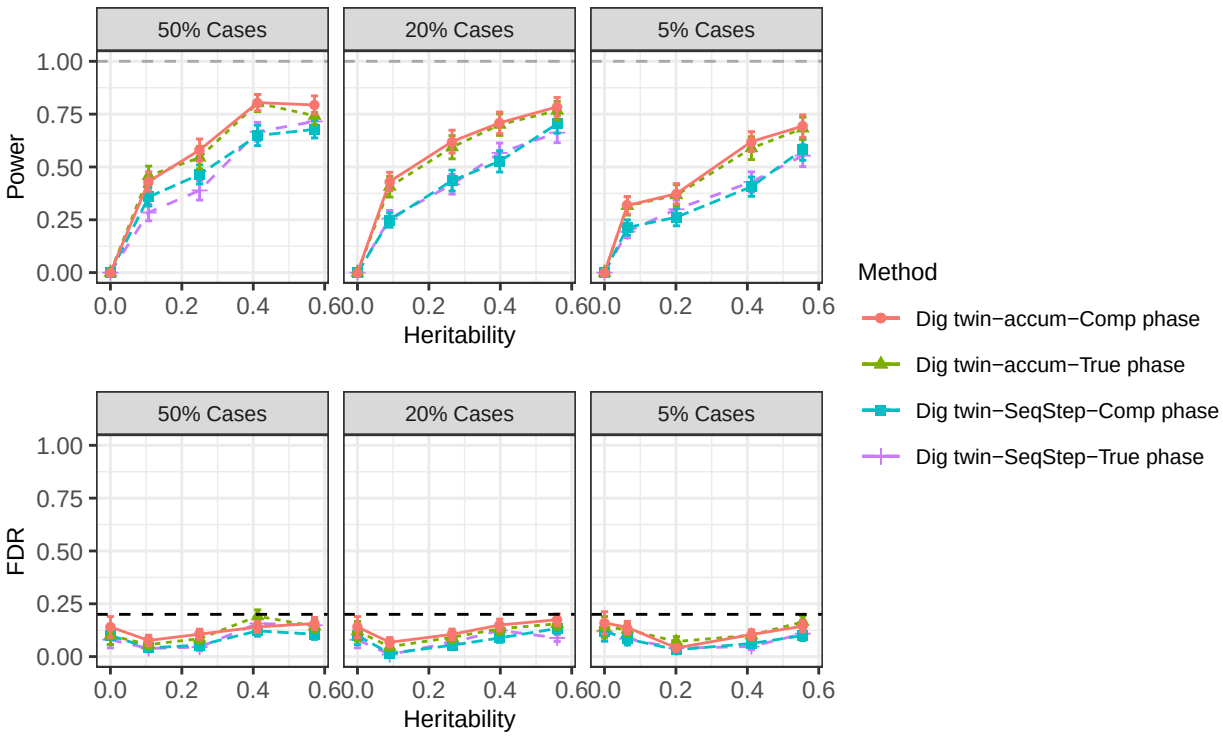

Figure S.4: A comparison of the Local Digital Twin Test with perfect phasing to that with computational phasing. The solid grey line indicates the nominal FDR threshold.

G.4 Full-genome simulation with a continuous response

We carry out a simulation study analogous to that of Section 5.2, now with a continuous trait. As a benchmark, we substitute the QFAM family-based association test [78, 79, 80, 81], for the TDT, but otherwise proceed in an identical manner. We report the results in Figure S.5. In this case, the Digital Twin Test has much lower power than the QFAM-based benchmarks, except for high values of heritability, and all procedures empirically control the FDR.

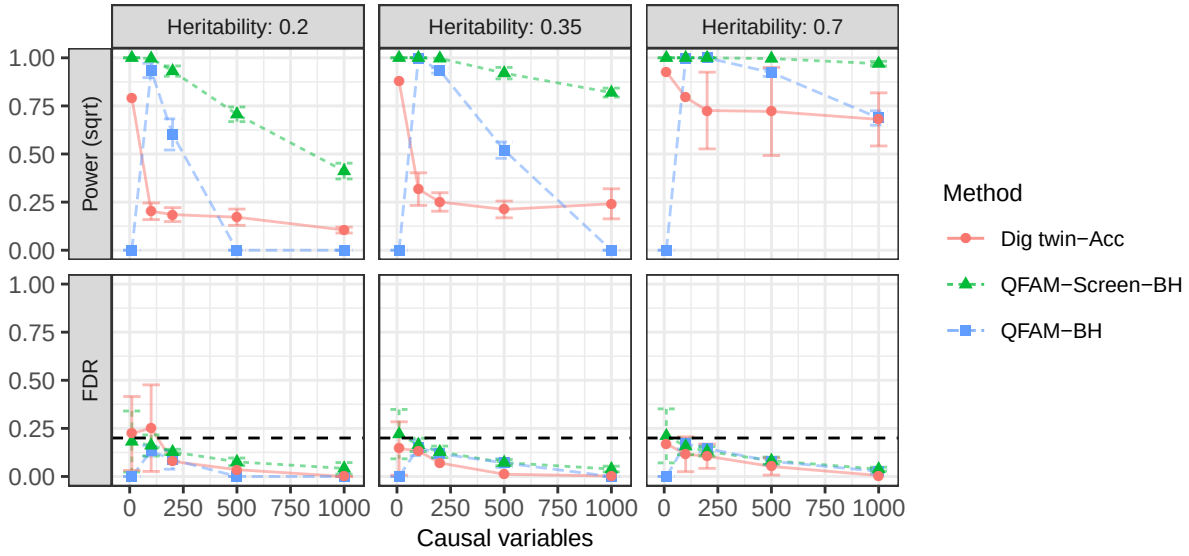

Figure S.5: FDR and power of the Local Digital Twin Test in a large full-genome simulation with a continuous trait, compared to the performance of the QFAM procedure. Other details as in Figure 7.

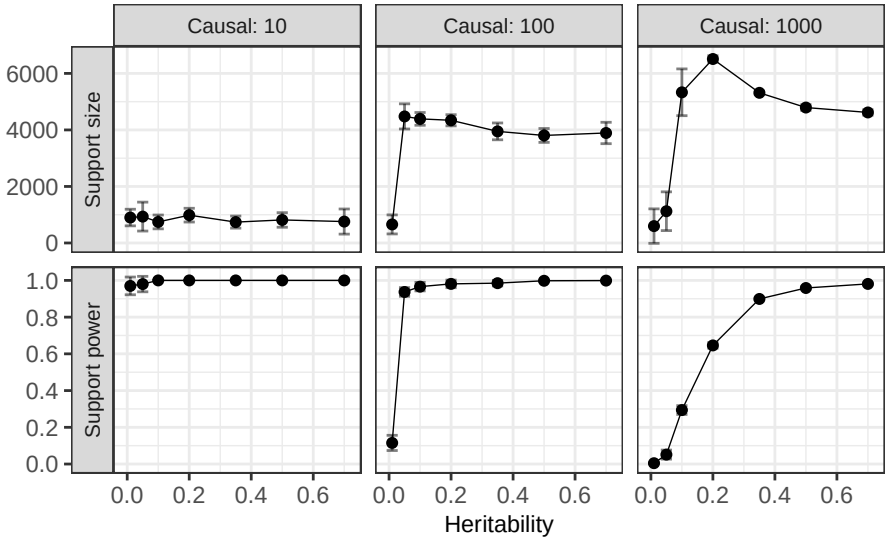

Figure S.6: Properties of the regression coefficients $\hat{\beta}$ in the full-genome simulations with continuous trait. Other details as in Figure S.2.