



DIPARTIMENTO DI  
INFORMATICA, SISTEMISTICA E  
COMUNICAZIONE

# Causal Inference in Genetic Trio Studies

*S. Bates, M. Sesia, C. Sabatti and E. Candès*

Stanford University

*Team: Daniele Maria Papetti, Alessandro Tundo, Matteo Vaghi*

*Causal Networks PhD Course 2021*

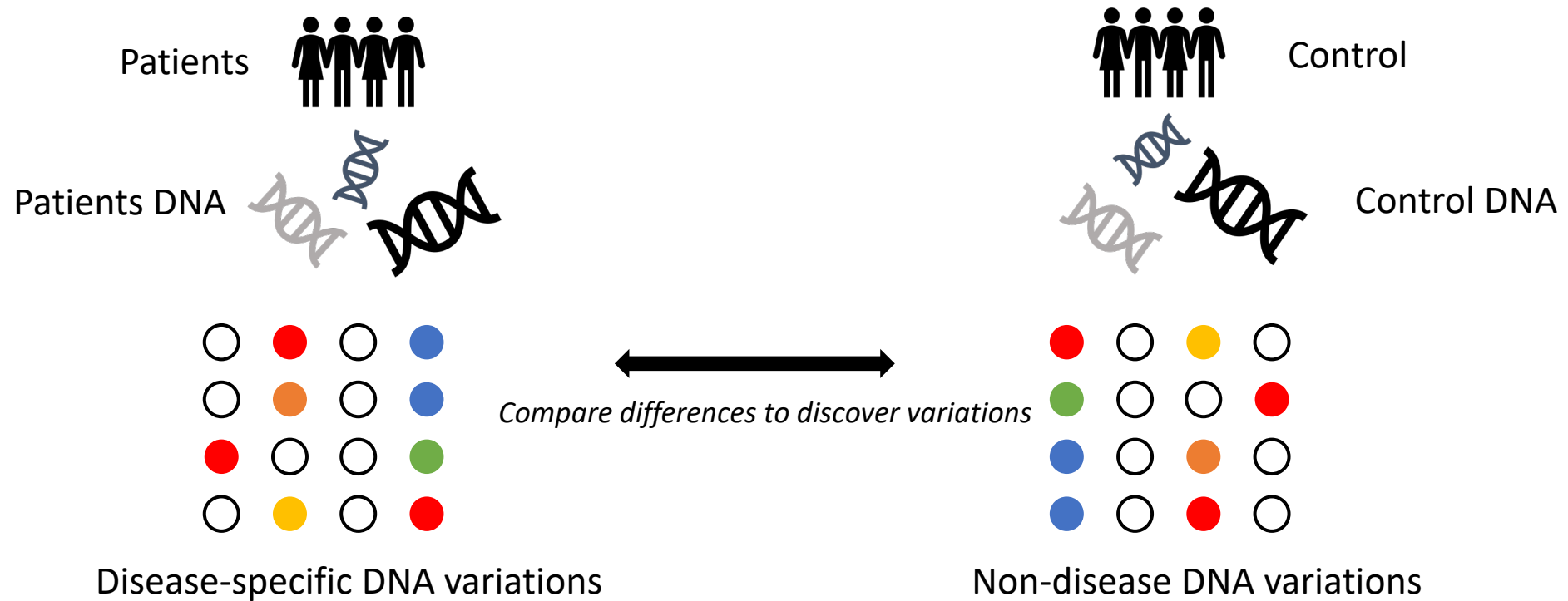
# Outline



- Background knowledge
- Main contribution
- Causality in trio design
- The Digital Twin Test (DTT): full-chromosome and local
- Case study: Autism Spectrum Disorder (ASD)
- Final remarks

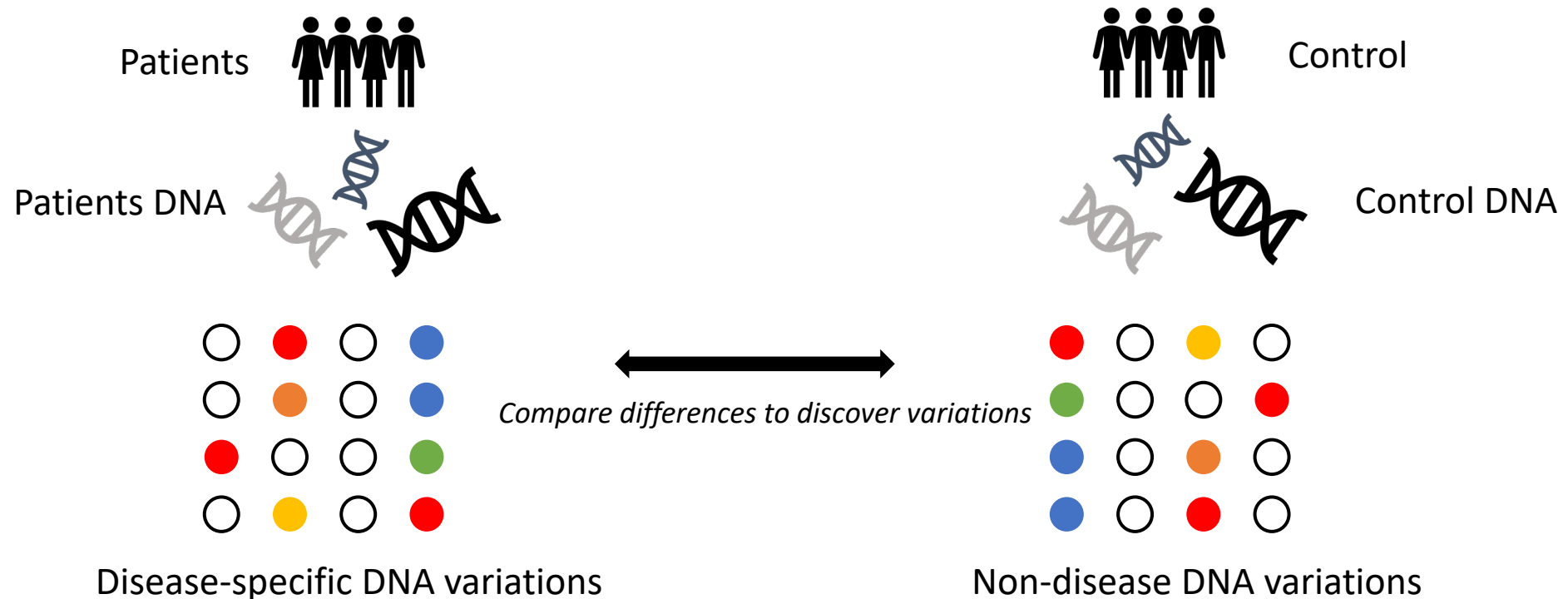
# Genome-wide Association Studies (GWAS)

**GWAS** discover regions of the genome containing *variants* that causally affect a *phenotype*, that is, identifying meaningful relationships between *genotypes* and *outcomes* of interest



# Genome-wide Association Studies (GWAS)

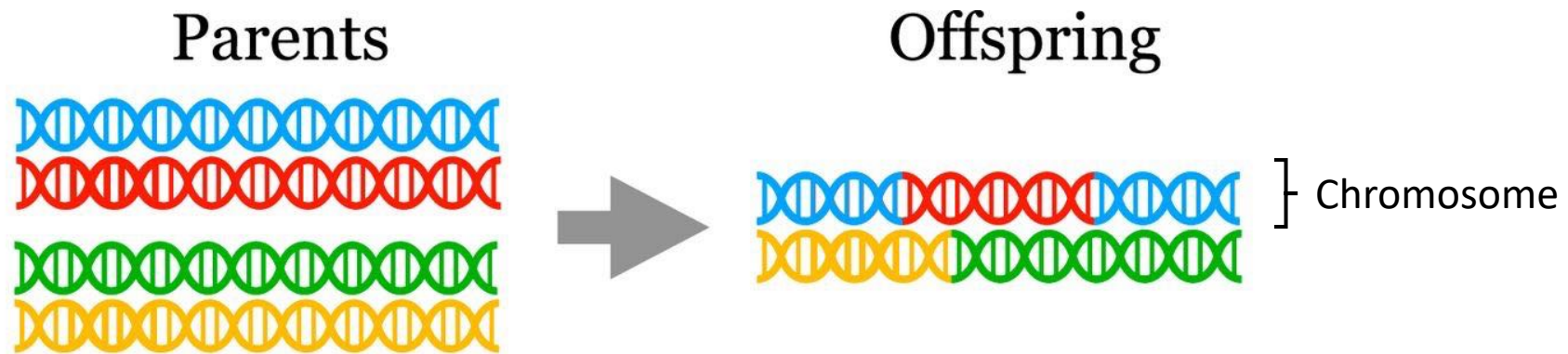
**GWAS** discover regions of the genome containing *variants* that causally affect a *phenotype*, that is, identifying meaningful relationships between *genotypes* and *outcomes* of interest



*N.B.: all true statistical associations represent relevant biological activity, irrelevant but true associations can arise from the confounding effect of environmental conditions or other factors*

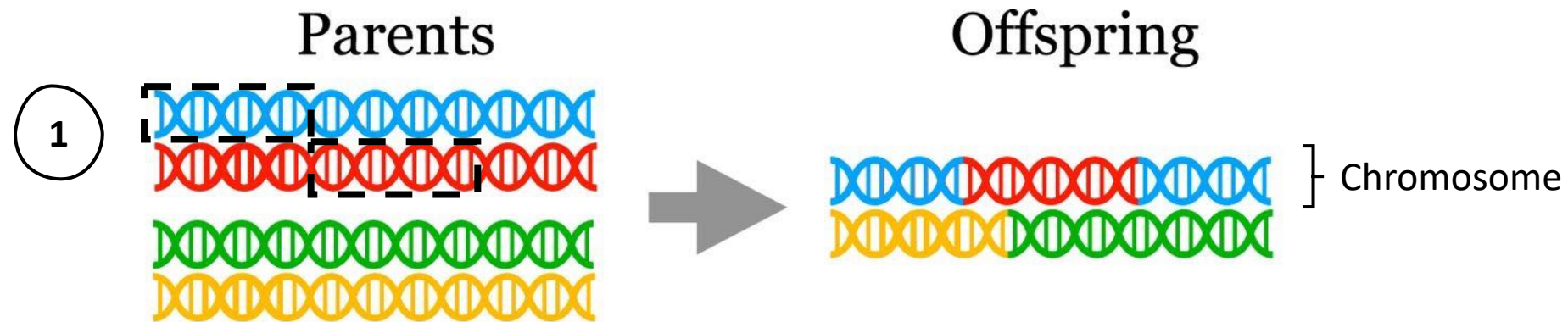
# Background Knowledge: Meiosis

It is a special type of cell division in sexually-reproducing organisms



# Background Knowledge: Meiosis

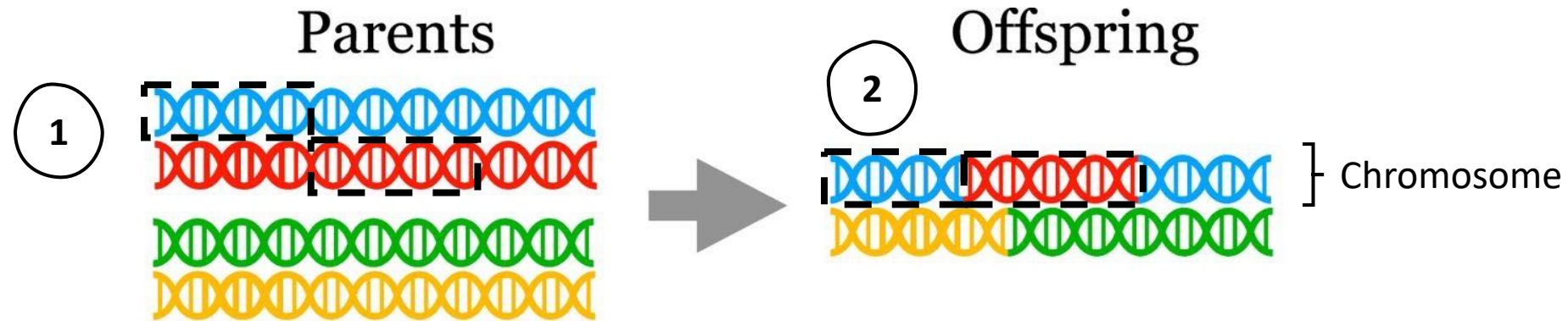
It is a special type of cell division in sexually-reproducing organisms



- 1 Prior to the division, genetic material from the paternal and maternal copies of each chromosome is crossed over

# Background Knowledge: Meiosis

It is a special type of cell division in sexually-reproducing organisms

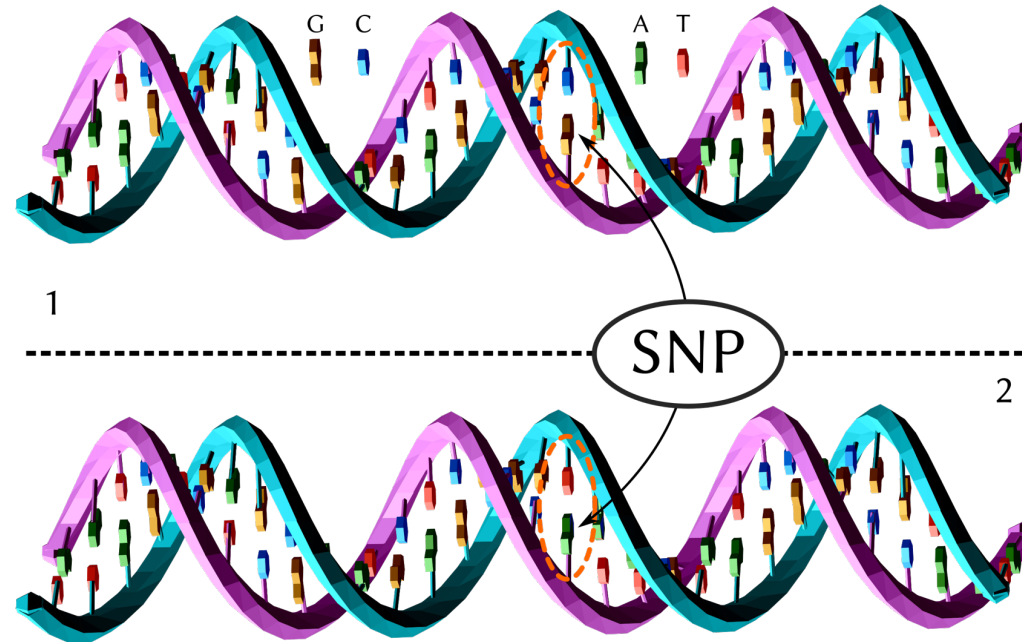


1 Prior to the division, genetic material from the paternal and maternal copies of each chromosome is crossed over

2 Creates new combinations of code on each chromosome

# Background Knowledge: Single-nucleotide Polymorphisms (SNPs)

SNPs are sites on the genome where 2 possible alleles occur in the population



**Allele:** one of the variants of a gene

**Haplotype:** set of observed alleles for an entire strand

*N.B.: SNPs on the same chromosome are dependent*



# Main Contribution

1

**Establishing causality in the trio design**

*formalization of family studies immunity w.r.t. population structure*

# Main Contribution

- 1 Establishing causality in the trio design**  
*formalization of family studies immunity w.r.t. population structure*
- 2 Identifying distinct causal regions**  
*localization of causal variants within windows in the full genome*

# Main Contribution

- 1 Establishing causality in the trio design**  
*formalization of family studies immunity w.r.t. population structure*
- 2 Identifying distinct causal regions**  
*localization of causal variants within windows in the full genome*
- 3 Testing multiple hypothesis**  
*creation of independent  $p$ -values for distinct regions without conservative corrections*

# Main Contribution

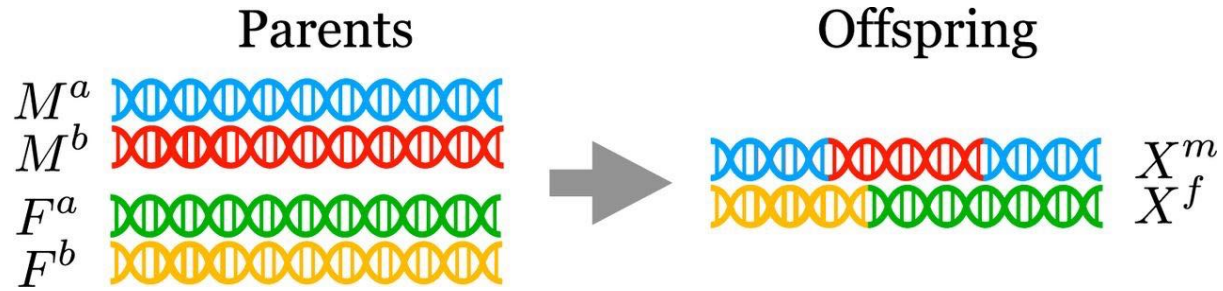
- 1 Establishing causality in the trio design**  
*formalization of family studies immunity w.r.t. population structure*
- 2 Identifying distinct causal regions**  
*localization of causal variants within windows in the full genome*
- 3 Testing multiple hypothesis**  
*creation of independent  $p$ -values for distinct regions without conservative corrections*
- 4 Leveraging black-box models and subject matter knowledge**  
*possibility of exploiting several multivariate models and domain information to increase power*

# Main Contribution

- 1 Establishing causality in the trio design**  
*formalization of family studies immunity w.r.t. population structure*
- 2 Identifying distinct causal regions**  
*localization of causal variants within windows in the full genome*
- 3 Testing multiple hypothesis**  
*creation of independent p-values for distinct regions without conservative corrections*
- 4 Leveraging black-box models and subject matter knowledge**  
*possibility of exploiting several multivariate models and domain information to increase power*

**Digital Twin Test:** *a method for finding causal regions that are immune to confounding variables*

# Notation

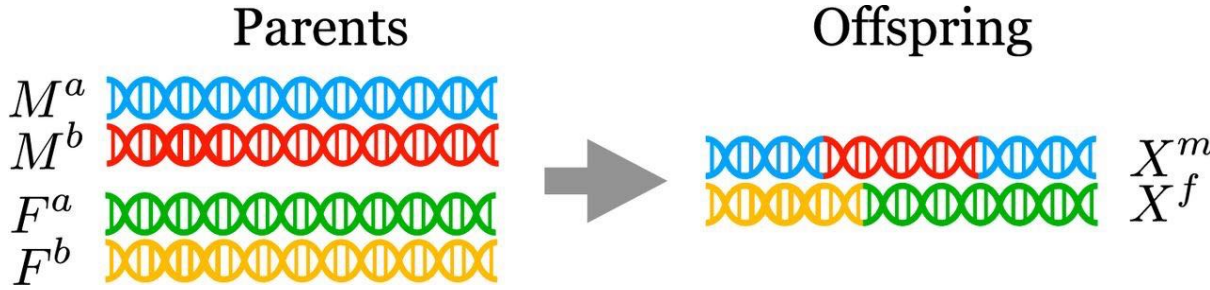


**Subjects:**  $(X_1^m, \dots, X_p^m) \in \{0,1\}^{n \times p}$ ,  $(X_1^f, \dots, X_p^f) \in \{0,1\}^{n \times p}$ ;

**Mothers:**  $(M_1^a, \dots, M_p^a) \in \{0,1\}^{n \times p}$ ,  $(M_1^b, \dots, M_p^b) \in \{0,1\}^{n \times p}$ ;

**Fathers:**  $(F_1^a, \dots, F_p^a) \in \{0,1\}^{n \times p}$ ,  $(F_1^b, \dots, F_p^b) \in \{0,1\}^{n \times p}$ .

# Notation



**Subjects:**  $(X_1^m, \dots, X_p^m) \in \{0,1\}^{n \times p}$ ,  $(X_1^f, \dots, X_p^f) \in \{0,1\}^{n \times p}$ ;

**Mothers:**  $(M_1^a, \dots, M_p^a) \in \{0,1\}^{n \times p}$ ,  $(M_1^b, \dots, M_p^b) \in \{0,1\}^{n \times p}$ ;

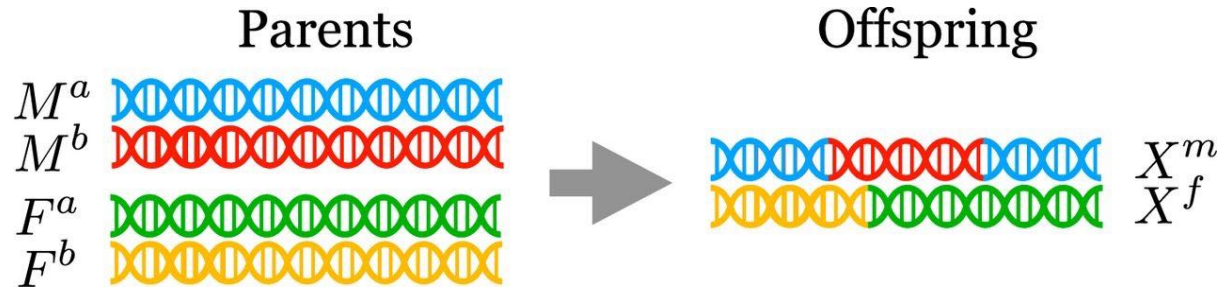
**Fathers:**  $(F_1^a, \dots, F_p^a) \in \{0,1\}^{n \times p}$ ,  $(F_1^b, \dots, F_p^b) \in \{0,1\}^{n \times p}$ .

**Offspring *genotypes* matrix:**  $X = X^m + X^f$



$X_j$ :  $j$ -th column of  $X$  representing the  $j$ -th genome site  
 $X^{(i)}$ :  $i$ -th row of  $X$  representing the subject  $i$

# Notation



**Subjects:**  $(X_1^m, \dots, X_p^m) \in \{0,1\}^{n \times p}$ ,  $(X_1^f, \dots, X_p^f) \in \{0,1\}^{n \times p}$ ;

**Mothers:**  $(M_1^a, \dots, M_p^a) \in \{0,1\}^{n \times p}$ ,  $(M_1^b, \dots, M_p^b) \in \{0,1\}^{n \times p}$ ;

**Fathers:**  $(F_1^a, \dots, F_p^a) \in \{0,1\}^{n \times p}$ ,  $(F_1^b, \dots, F_p^b) \in \{0,1\}^{n \times p}$ .

**Offspring *genotypes* matrix:**  $X = X^m + X^f$



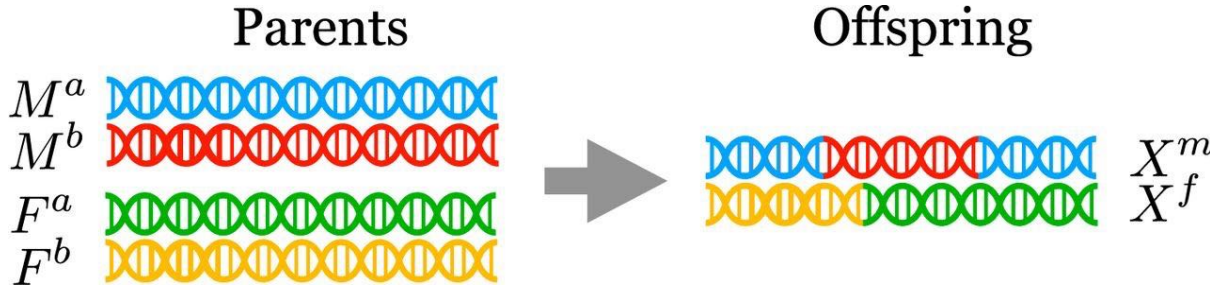
$X_j$ :  $j$ -th column of  $X$  representing the  $j$ -th genome site

$X^{(i)}$ :  $i$ -th row of  $X$  representing the subject  $i$

**Ancestral *haplotypes*:**  $A = (M^a, M^b, F^a, F^b)$



# Notation



**Subjects:**  $(X_1^m, \dots, X_p^m) \in \{0,1\}^{n \times p}$ ,  $(X_1^f, \dots, X_p^f) \in \{0,1\}^{n \times p}$ ;

**Mothers:**  $(M_1^a, \dots, M_p^a) \in \{0,1\}^{n \times p}$ ,  $(M_1^b, \dots, M_p^b) \in \{0,1\}^{n \times p}$ ;

**Fathers:**  $(F_1^a, \dots, F_p^a) \in \{0,1\}^{n \times p}$ ,  $(F_1^b, \dots, F_p^b) \in \{0,1\}^{n \times p}$ .

**Offspring genotypes matrix:**  $X = X^m + X^f$    $X_j$ :  $j$ -th column of  $X$  representing the  $j$ -th genome site  
 $X^{(i)}$ :  $i$ -th row of  $X$  representing the subject  $i$

**Ancestral haplotypes:**  $A = (M^a, M^b, F^a, F^b)$

**SNP  $X_j^m$ :** single-nucleotide polymorphism inherited either from  $M_j^a$  or  $M_j^b$  with equal probability

# Causality in Trio Design: Establishing Causality

## Main ideas

- 1 gene inheritance process can be seen as a *high-dimensional randomized experiment*

# Causality in Trio Design: Establishing Causality

## Main ideas

- 1 gene inheritance process can be seen as a *high-dimensional randomized experiment*
- 2 conditioning on the parental *haplotypes* removes external possible *confounders* from the meiosis process

# Causality in Trio Design: Establishing Causality

## Main ideas

- 1 gene inheritance process can be seen as a *high-dimensional randomized experiment*
- 2 conditioning on the parental *haplotypes* removes external possible *confounders* from the meiosis process

$H_0: X_j \perp\!\!\!\perp Y \mid A$  **Null hypothesis** that a *SNP*  $X_j$  is independent of the response  $Y$  after conditioning on the parental haplotypes  $A$

# Causality in Trio Design: Establishing Causality

## Main ideas

- 1 gene inheritance process can be seen as a *high-dimensional randomized experiment*
- 2 conditioning on the parental *haplotypes* removes external possible *confounders* from the meiosis process

$H_0: X_j \perp\!\!\!\perp Y \mid A$  **Null hypothesis** that a *SNP*  $X_j$  is independent of the response  $Y$  after conditioning on the parental haplotypes  $A$

**External confounder**  $X \mid (A, Z = z) \stackrel{\text{def}}{=} X \mid (A, Z = z')$  for any  $z$  and  $z'$

# Causality in Trio Design: Establishing Causality

## Main ideas

- 1 gene inheritance process can be seen as a *high-dimensional randomized experiment*
- 2 conditioning on the parental *haplotypes* removes external possible *confounders* from the meiosis process

$H_0: X_j \perp\!\!\!\perp Y \mid A$  **Null hypothesis** that a *SNP*  $X_j$  is independent of the response  $Y$  after conditioning on the parental haplotypes  $A$

**External confounder**  $X \mid (A, Z = z) \stackrel{\text{def}}{=} X \mid (A, Z = z')$  for any  $z$  and  $z'$

↓

$Z \perp\!\!\!\perp X_j \mid A$  *The confounder  $Z$  is independent of the SNP  $j$  given  $A$*

# Causality in Trio Design: Establishing Causality

## Main ideas

- 1 gene inheritance process can be seen as a *high-dimensional randomized experiment*
- 2 conditioning on the parental *haplotypes* removes external possible *confounders* from the meiosis process

$H_0: X_j \perp\!\!\!\perp Y \mid A$  **Null hypothesis** that a *SNP*  $X_j$  is independent of the response  $Y$  after conditioning on the parental haplotypes  $A$

**External confounder**  $X \mid (A, Z = z) \stackrel{\text{def}}{=} X \mid (A, Z = z')$  for any  $z$  and  $z'$

$\Downarrow$

$Z \perp\!\!\!\perp X_j \mid A$  The confounder  $Z$  is independent of the *SNP*  $j$  given  $A$

$\Downarrow$

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

If  $X$  and  $Y$  are associated after conditioning on  $A$ , the association is not due to confounder  $Z$

# Causality in Trio Design: Establishing Causality

Let  $Z$  be an external confounder, then any valid test of the null hypothesis  $H_0$  is also a valid test of the stronger null hypothesis that accounts for the confounder  $Z$ :

$$H'_0: Y \perp\!\!\!\perp X_j \mid (A, Z)$$



# Causality in Trio Design: Establishing Causality

Let  $Z$  be an external confounder, then any valid test of the null hypothesis  $H_0$  is also a valid test of the stronger null hypothesis that accounts for the confounder  $Z$ :

$$H'_0: Y \perp\!\!\!\perp X_j \mid (A, Z)$$

**Note 1:** if we reject  $H_0$ , the dependence between  $X_j$  and  $Y$  cannot be due to an external confounder  $Z$

# Causality in Trio Design: Establishing Causality

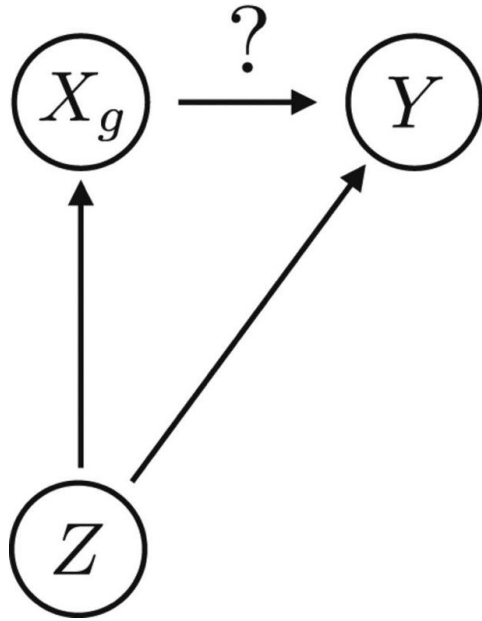
Let  $Z$  be an external confounder, then any valid test of the null hypothesis  $H_0$  is also a valid test of the stronger null hypothesis that accounts for the confounder  $Z$ :

$$H'_0: Y \perp\!\!\!\perp X_j \mid (A, Z)$$

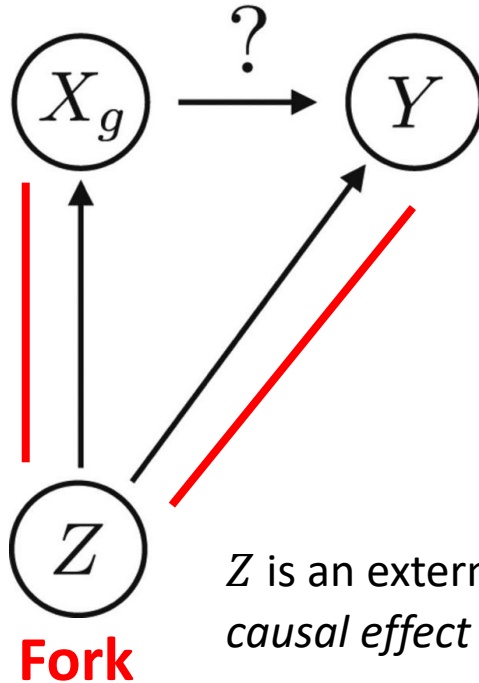
**Note 1:** if we reject  $H_0$ , the dependence between  $X_j$  and  $Y$  cannot be due to an external confounder  $Z$

**Note 2:** if we reject  $H_0$ , it does not yet imply that  $X_j$  is the causal *SNP*, but it implies that there is an association on the chromosome that is not the result of external confounding

# Graphical Causal Model

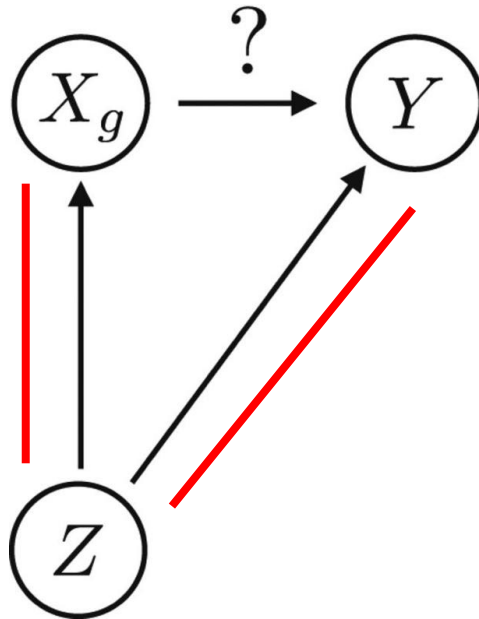


# Graphical Causal Model

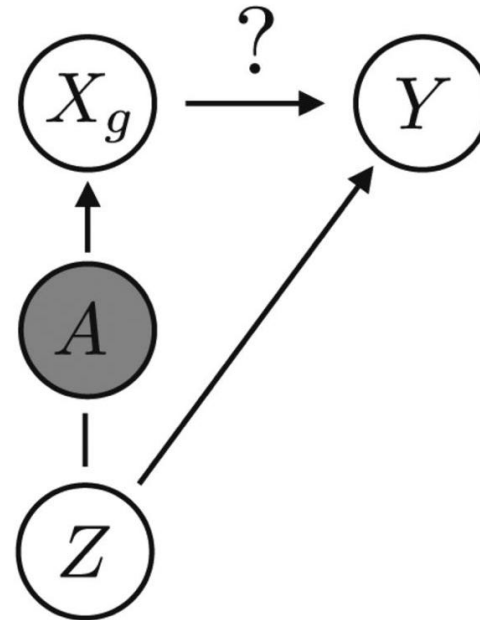


$Z$  is an external confounder that can create an association between  $X_g$  and  $Y$  *even if there is no causal effect* (due to the **fork** structure)

# Graphical Causal Model



**Fork**

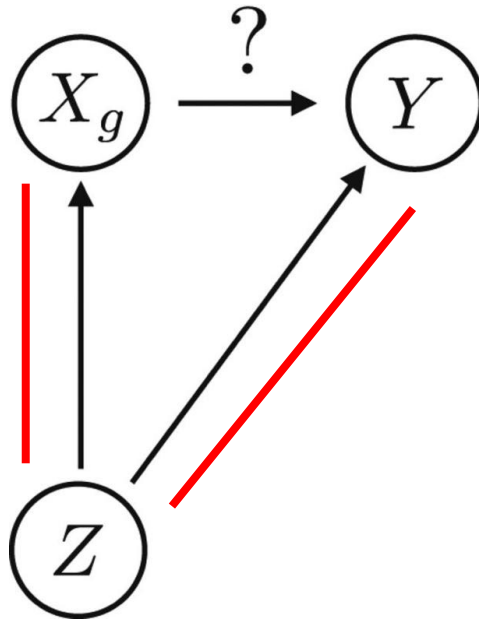


Structural Equation Model M:  $(A, Z) = f_{AZ}(N_{AZ}), \quad X = f_X(A, N_X), \quad Y = f_Y(X, Z, N_Y)$

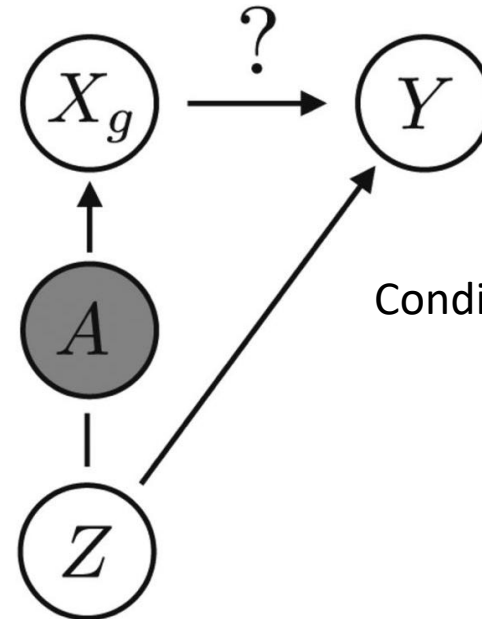
$f_{AZ}, f_X, f_Y$  are fixed functions

$N_{AZ}, N_X, N_Y$  are the exogenous variables

# Graphical Causal Model



**Fork**



Conditioning on  $A$  makes  $Z$  independent of  $X_g$

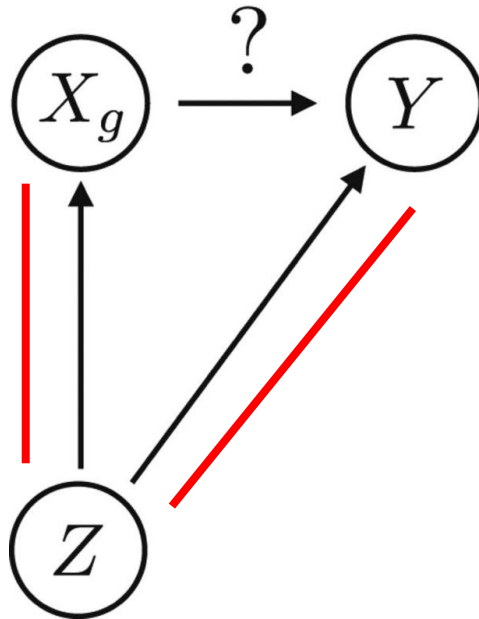
**Chain**

Structural Equation Model  $M$ :  $(A, Z) = f_{AZ}(N_{AZ}), \quad X = f_X(A, N_X), \quad Y = f_Y(X, Z, N_Y)$

$f_{AZ}, f_X, f_Y$  are fixed functions

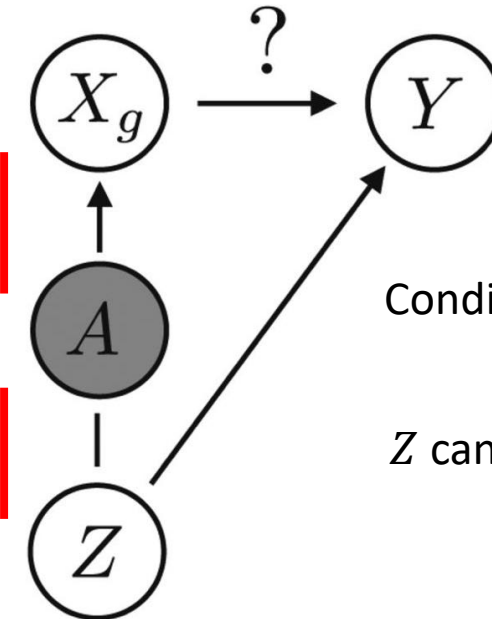
$N_{AZ}, N_X, N_Y$  are the exogenous variables

# Graphical Causal Model



**Fork**

**Chain**



Conditioning on  $A$  makes  $Z$  independent of  $X_g$



$Z$  cannot be responsible for the remaining association between  $X_g$  and  $Y$

Structural Equation Model M:  $(A, Z) = f_{AZ}(N_{AZ}), \quad X = f_X(A, N_X), \quad Y = f_Y(X, Z, N_Y)$

$f_{AZ}, f_X, f_Y$  are fixed functions

$N_{AZ}, N_X, N_Y$  are the exogenous variables

# Discussion of Possible Confounders

(virtually) all *confounders* in genetic studies do not affect the transmission of the genetic information



thus *external confounders* which are correctly accounted for in the trio design



# Discussion of Possible Confounders

## Examples of external confounders



Environmental conditions after conception



Population structure, ethnicity, location



Cryptic relatedness



Family effects, altruistic genes



Assortative mating

# Discussion of Possible Confounders

## Examples of external confounders



Environmental conditions after conception



Population structure, ethnicity, location



Cryptic relatedness



Family effects, altruistic genes



Assortative mating

## Examples of not external confounders



Germline mutations



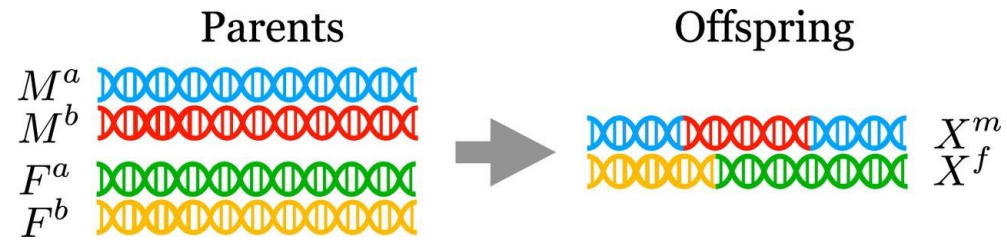
Unmeasured SNPs

# The Randomness in Inheritance

The process by which subject's two haplotypes arise from the parental haplotypes is modelled as a **hidden Markov model (HMM)**

# The Randomness in Inheritance

The process by which subject's two haplotypes arise from the parental haplotypes is modelled as a **hidden Markov model (HMM)**

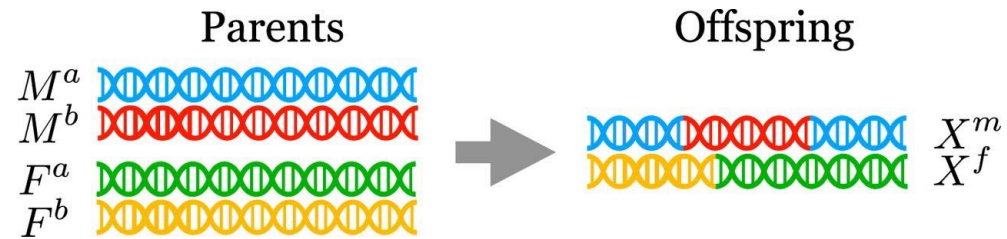


Model for a single observation on one chromosome (e.g.:  $X^m$ )

$$X_j^m = \begin{cases} a & \text{if site } j \text{ is copied from } M^a \\ b & \text{if site } j \text{ is copied from } M^b \end{cases}$$

# The Randomness in Inheritance

The process by which subject's two haplotypes arise from the parental haplotypes is modelled as a **hidden Markov model (HMM)**



Model for a single observation on one chromosome (e.g.:  $X^m$ )

$$X_j^m = \begin{cases} a & \text{if site } j \text{ is copied from } M^a \\ b & \text{if site } j \text{ is copied from } M^b \end{cases}$$

$$P(X_1^m = a) = \frac{1}{2}$$



$$P(X_j^m = x_{j-1}^m | X_{1:(j-1)}^m = x_{1:(j-1)}^m) = \frac{1}{2} (1 + e^{-2d_j})$$

$d_j$  is the genetic distance between SNPs  $j - 1$  and  $j$

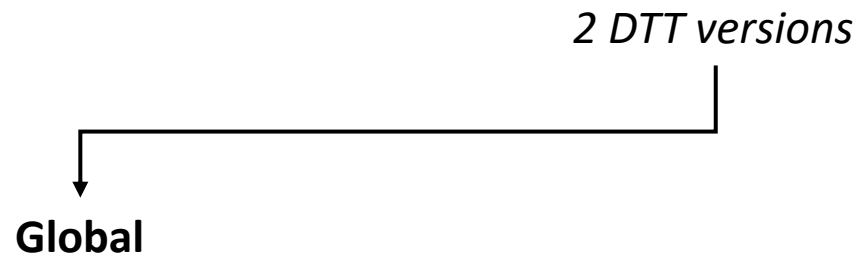
HMM describes the distribution of  $X^m$  given  $M^a$  and  $M^b$  ( $F^a, F^b$ )

# The Digital Twin Test (DTT)

## Goals

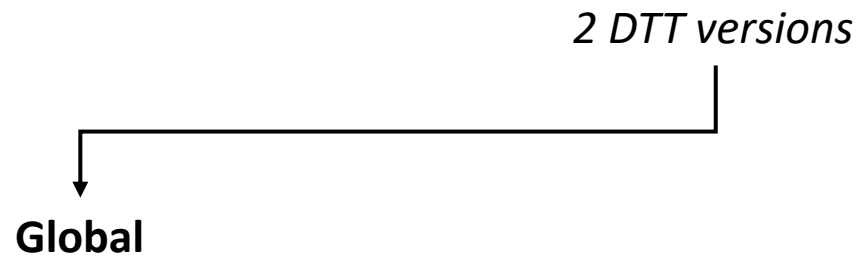
-  to determine whether a observed trait has any genetic basis, that is, to test the  $H_0: X_j \perp\!\!\!\perp Y \mid A$
-  to find regions of the genome that contain causal variants

# The Digital Twin Test (DTT)



to determine if the observed trait is caused  
by a set of SNPs on the chromosome

# The Digital Twin Test (DTT)

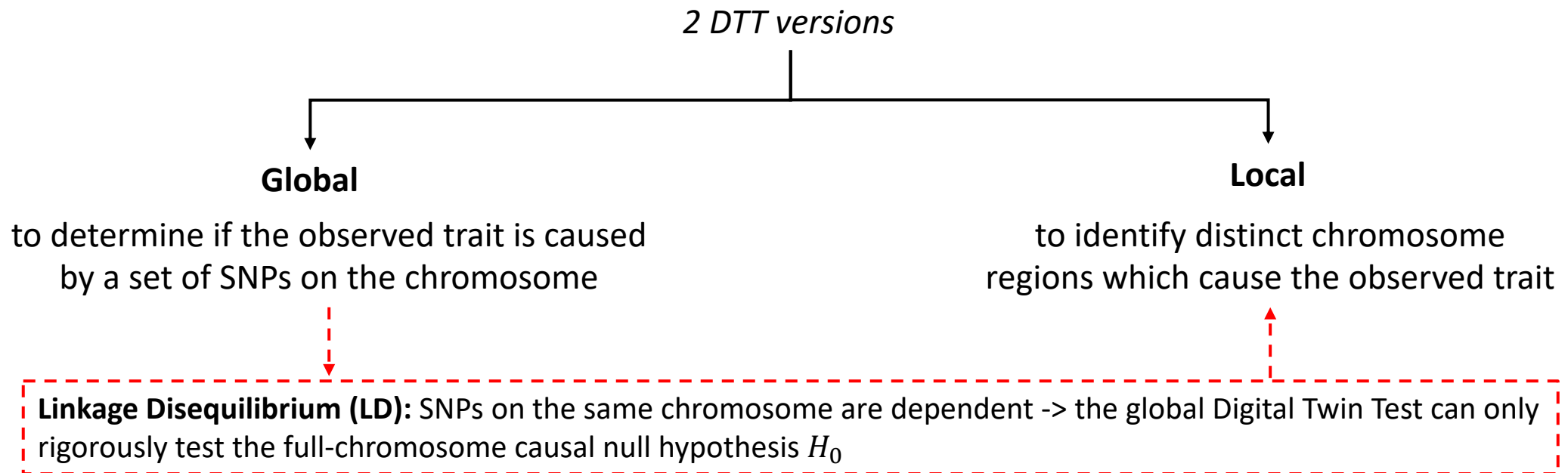


to determine if the observed trait is caused  
by a set of SNPs on the chromosome

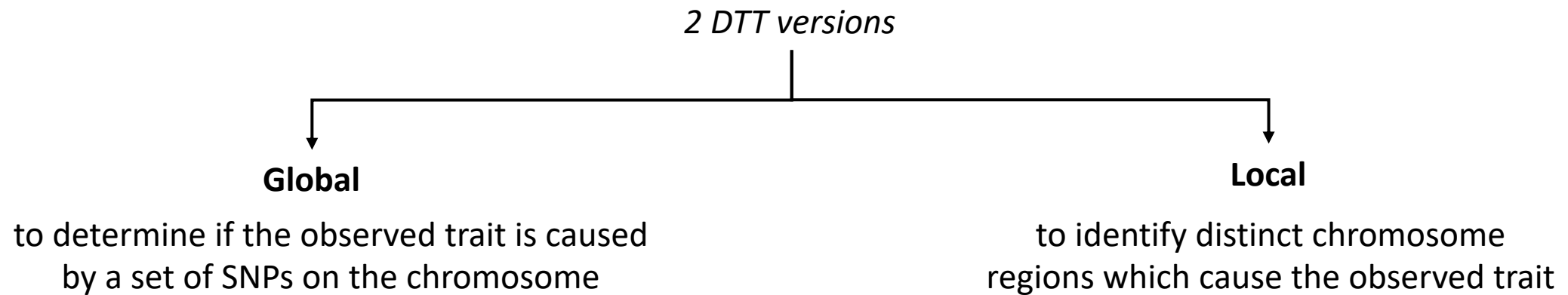
**Linkage Disequilibrium (LD):** SNPs on the same chromosome are dependent -> the global Digital Twin Test can only rigorously test the full-chromosome causal null hypothesis  $H_0$



# The Digital Twin Test (DTT)

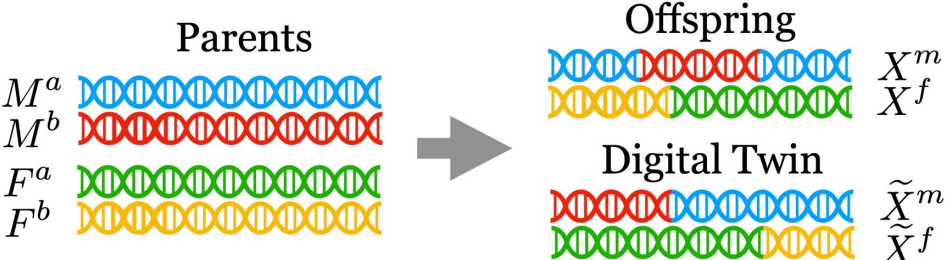


# The Digital Twin Test (DTT)



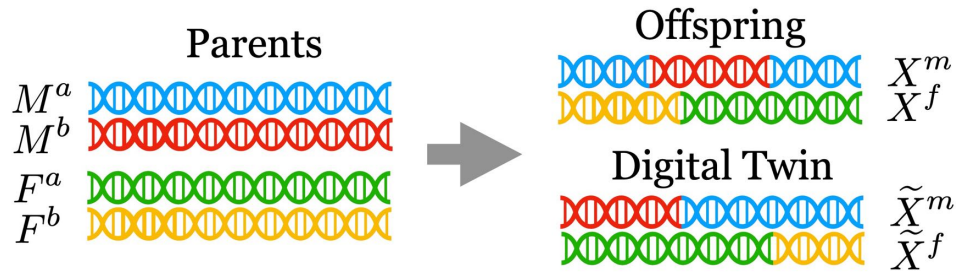
**N.B.:** *the DTT is a natural randomization test in a trio design because it replicates the random mechanism generating the data!*

# The full-chromosome Digital Twin Test



$$H_0: X_C \perp\!\!\!\perp Y \mid A$$

# The full-chromosome Digital Twin Test



Compute  $t^* = T((X_{-C}, X_C), Y)$

$H_0: X_C \perp\!\!\!\perp Y \mid A$

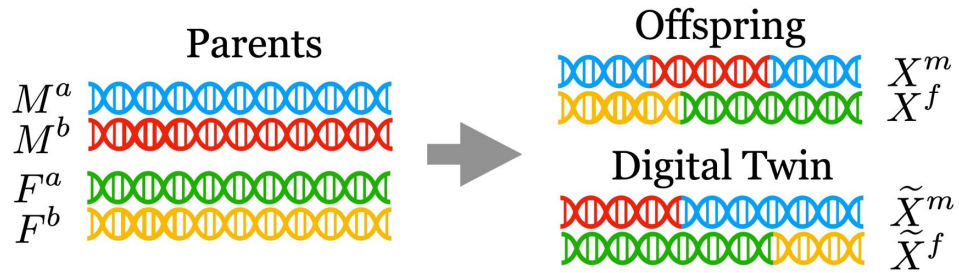
$C \subset \{1, \dots, p\}$  set of SNPs on chromosome

$X_C$  denotes  $(X_j)_{j \in C}$

$X_{-C}$  denotes  $(X_j)_{j \notin C}$

$T(\cdot)$  can be any statistic

# The full-chromosome Digital Twin Test



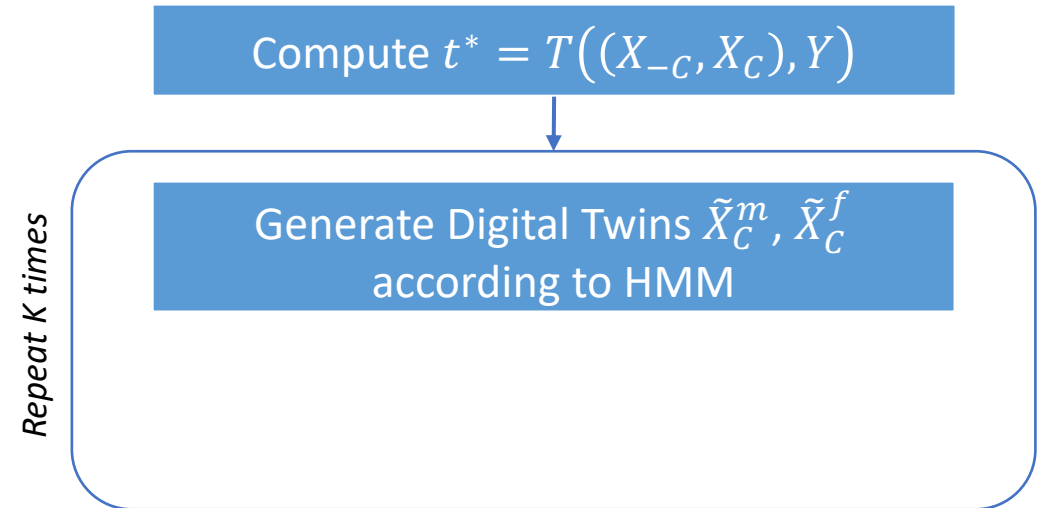
$$H_0: X_C \perp\!\!\!\perp Y \mid A$$

$C \subset \{1, \dots, p\}$  set of SNPs on chromosome

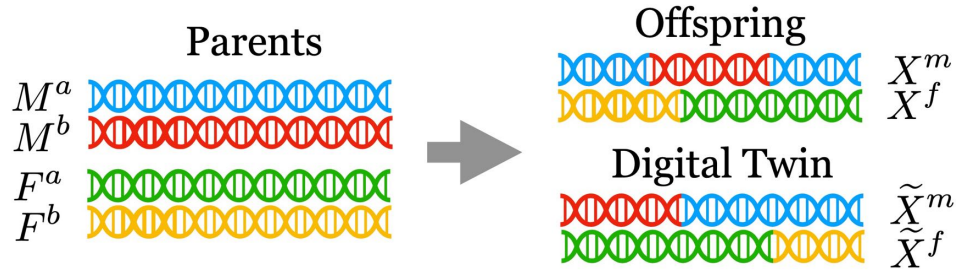
$X_C$  denotes  $(X_j)_{j \in C}$

$X_{-C}$  denotes  $(X_j)_{j \notin C}$

$T(\cdot)$  can be any statistic



# The full-chromosome Digital Twin Test



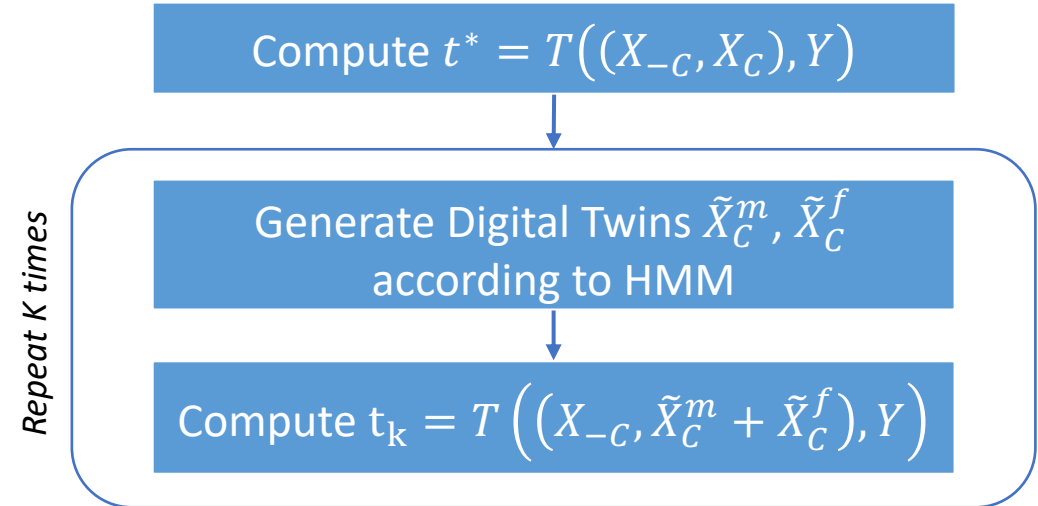
$$H_0: X_C \perp\!\!\!\perp Y \mid A$$

$C \subset \{1, \dots, p\}$  set of SNPs on chromosome

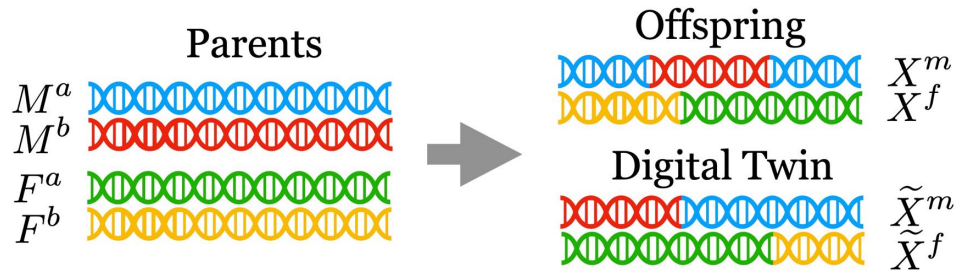
$X_C$  denotes  $(X_j)_{j \in C}$

$X_{-C}$  denotes  $(X_j)_{j \notin C}$

$T(\cdot)$  can be any statistic



# The full-chromosome Digital Twin Test



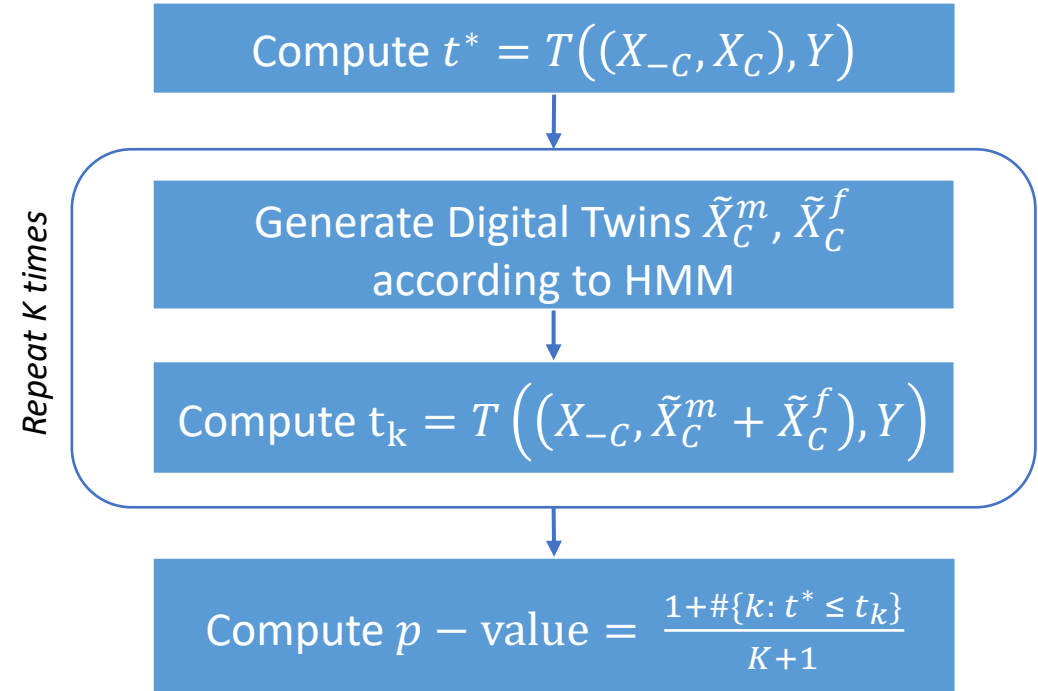
$$H_0: X_C \perp\!\!\!\perp Y \mid A$$

$C \subset \{1, \dots, p\}$  set of SNPs on chromosome

$X_C$  denotes  $(X_j)_{j \in C}$

$X_{-C}$  denotes  $(X_j)_{j \notin C}$

$T(\cdot)$  can be any statistic



The algorithm returns a  $p$ -value, and the corresponding level  $\alpha$  hypothesis test rejects when this  $p$ -value is less than  $\alpha$ .

# Full-chromosome DTT Evaluation

## **Experimental setting**

1. Parent-offspring generated by real haplotypes from UK Biobank dataset
2. Synthetic population of 2500 parent-offspring trios



# Full-chromosome DTT Evaluation

## **Experimental setting**

1. Parent-offspring generated by real haplotypes from UK Biobank dataset
2. Synthetic population of 2500 parent-offspring trios
3. Logistic regression model to generate parent-offspring trait

# Full-chromosome DTT Evaluation

## **Experimental setting**

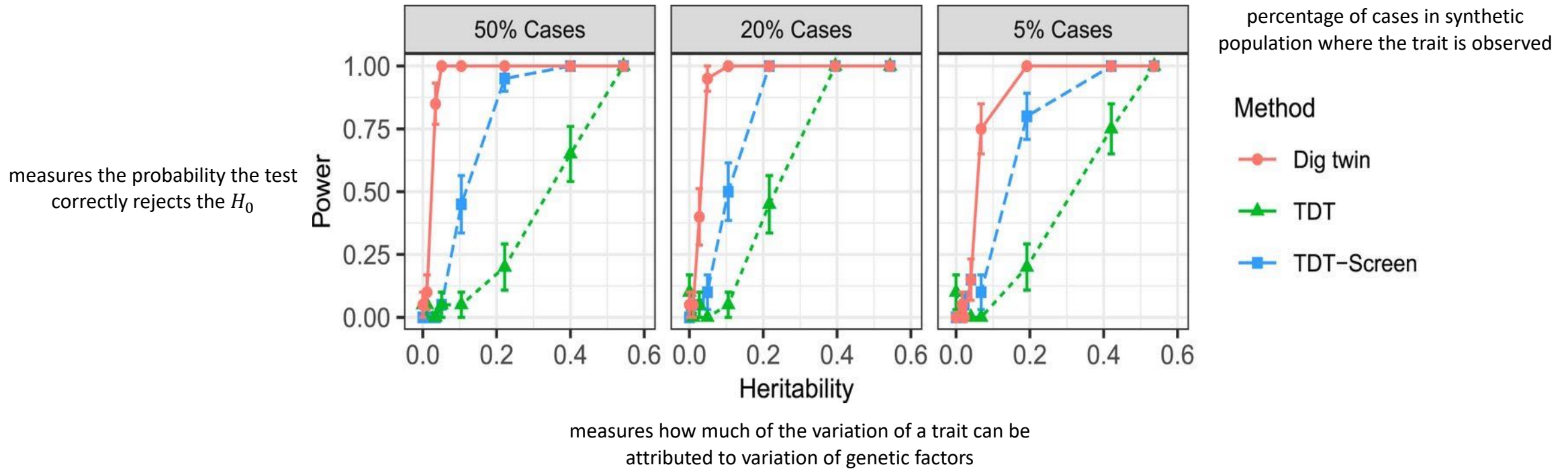
1. Parent-offspring generated by real haplotypes from UK Biobank dataset
2. Synthetic population of 2500 parent-offspring trios
3. Logistic regression model to generate parent-offspring trait
4. Confidence level  $\alpha = 0.05$

# Full-chromosome DTT Evaluation

## Experimental setting

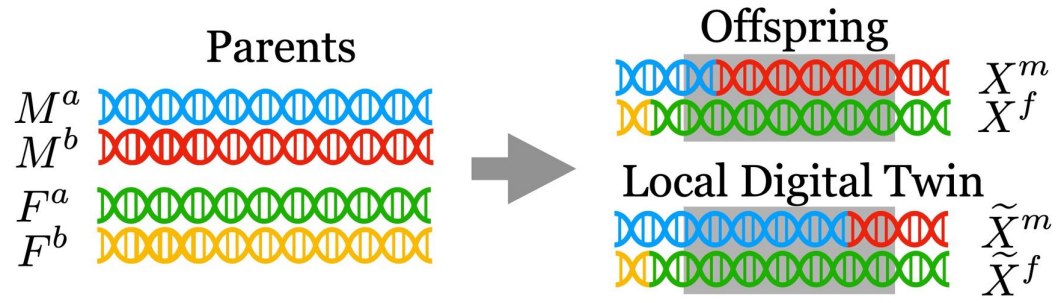
1. Parent-offspring generated by real haplotypes from UK Biobank dataset
2. Synthetic population of 2500 parent-offspring trios
3. Logistic regression model to generate parent-offspring trait
4. Confidence level  $\alpha = 0.05$
5. Comparison with TDT / TDT-Screen algorithms

# Full-chromosome DTT Evaluation



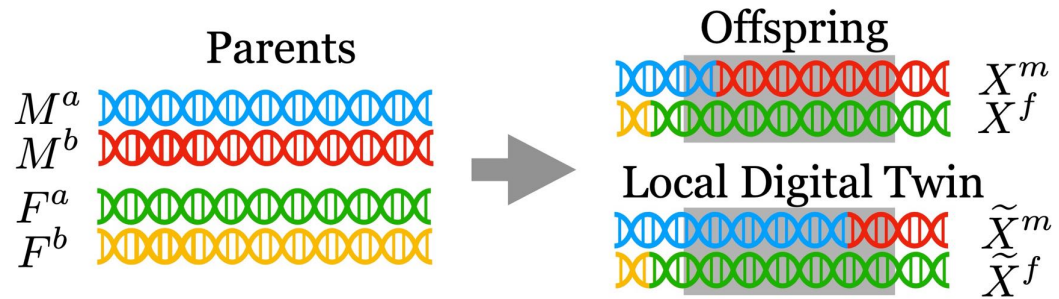
DTT outperforms the benchmark algorithms

# Local DTT



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

# Local DTT



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

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

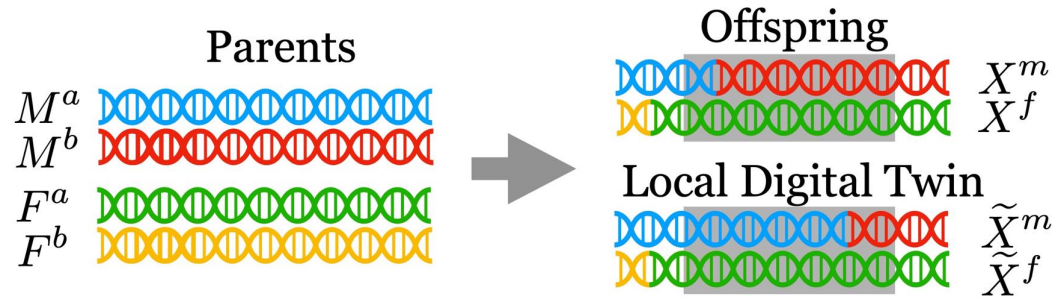
$G$  is a partition of  $\{1, \dots, p\}$ ,  $g \in G$  is group of SNPs

$X_g$  denotes  $(X_j)_{j \in g}$

$X_{-g}$  denotes  $(X_j)_{j \notin g}$

$T(\cdot)$  can be any statistic

# Local DTT



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

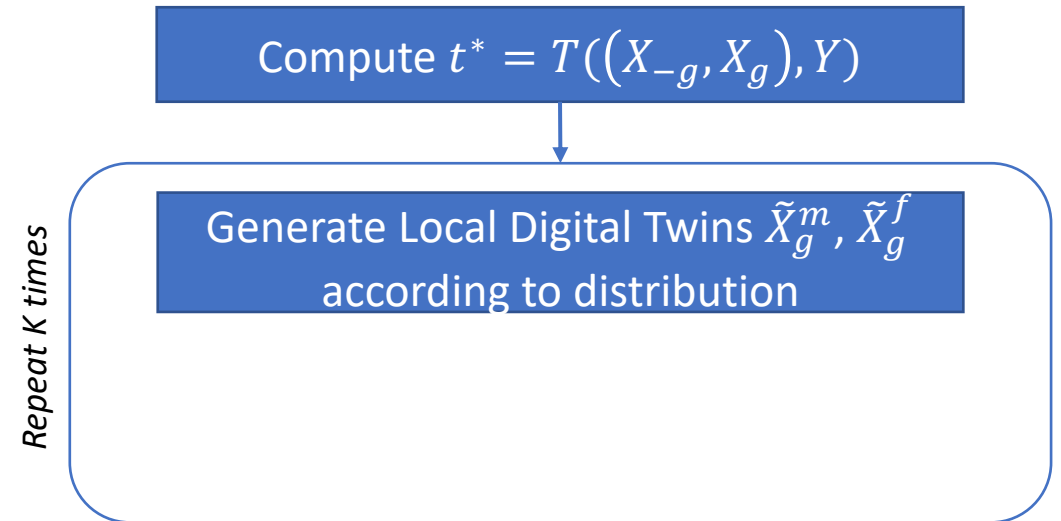
$G$  is a partition of  $\{1, \dots, p\}$ ,  $g \in G$  is group of SNPs

$X_g$  denotes  $(X_j)_{j \in g}$

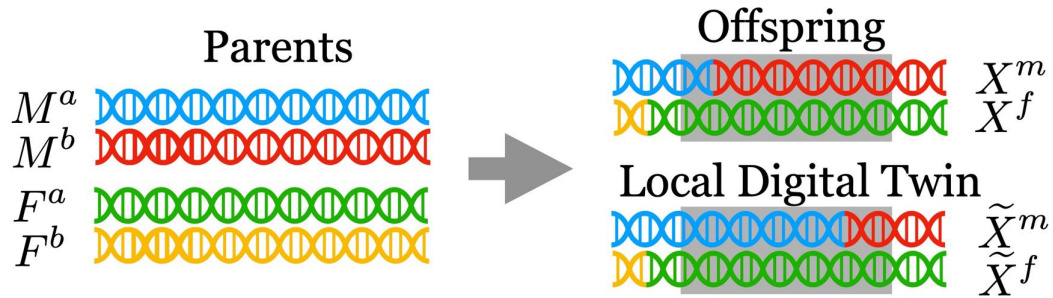
$X_{-g}$  denotes  $(X_j)_{j \notin g}$

$T(\cdot)$  can be any statistic

Sampling distribution of Local Digital Twins  $(X_g^m, X_g^f) \mid (X_{-g}^m, X_{-g}^f, A)$



# Local DTT



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

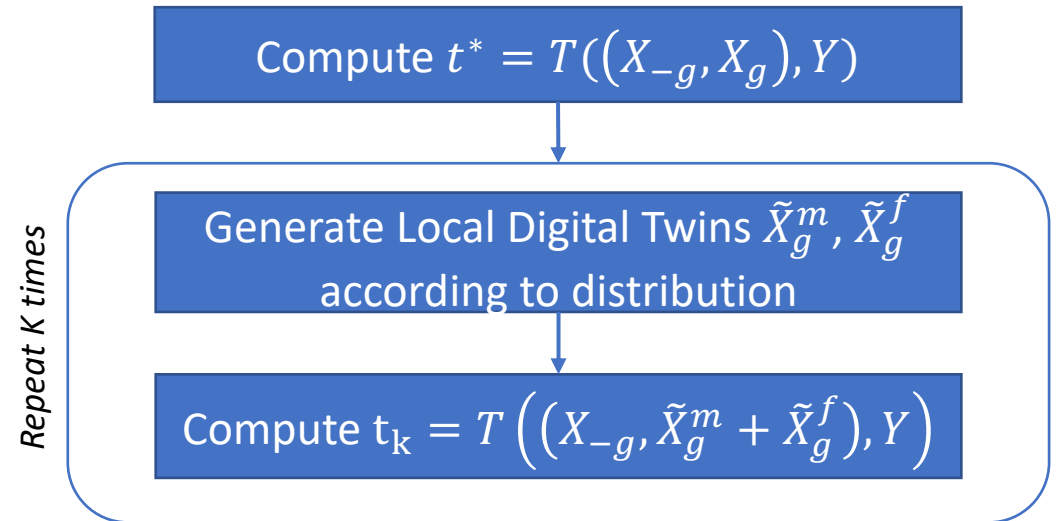
$G$  is a partition of  $\{1, \dots, p\}$ ,  $g \in G$  is group of SNPs

$X_g$  denotes  $(X_j)_{j \in g}$

$X_{-g}$  denotes  $(X_j)_{j \notin g}$

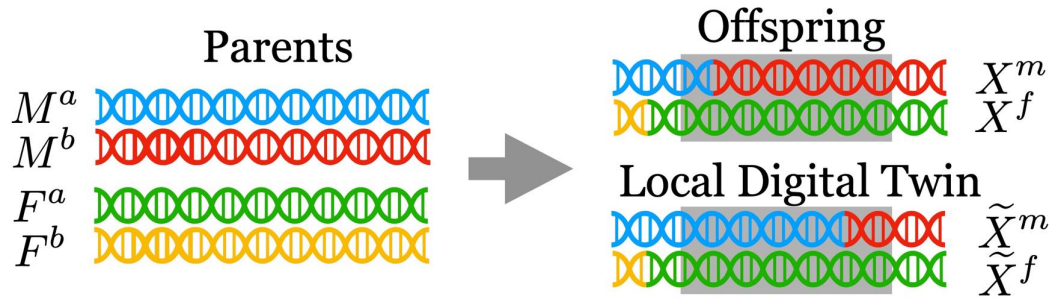
$T(\cdot)$  can be any statistic

Sampling distribution of Local Digital Twins  $(X_g^m, X_g^f) \mid (X_{-g}^m, X_{-g}^f, A)$





# Local DTT



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

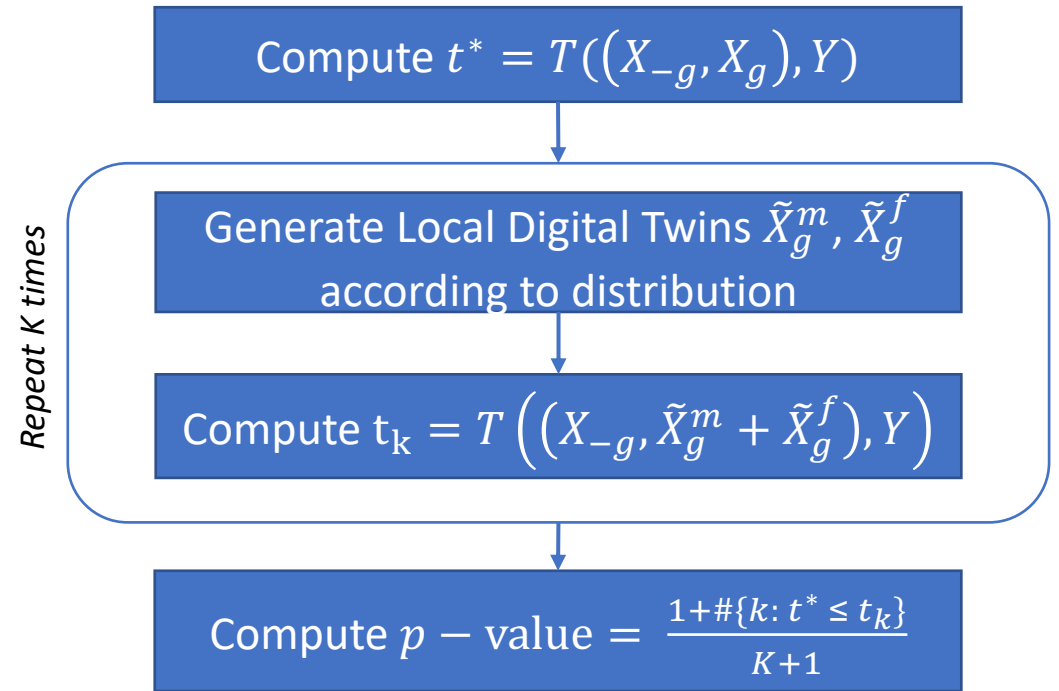
$G$  is a partition of  $\{1, \dots, p\}$ ,  $g \in G$  is group of SNPs

$X_g$  denotes  $(X_j)_{j \in g}$

$X_{-g}$  denotes  $(X_j)_{j \notin g}$

$T(\cdot)$  can be any statistic

Sampling distribution of Local Digital Twins  $(X_g^m, X_g^f) \mid (X_{-g}^m, X_{-g}^f, A)$



The algorithm returns  $p$ -values that require statistical corrections (e.g., Bonferroni or BH) to ensure independency

*(a slightly modification of the algorithm leads to independent  $p$ -values)*

# Local DTT Evaluation

## **Experimental setting**

1. Synthetic population of 10,000 parent-offspring trios

# Local DTT Evaluation

## **Experimental setting**

1. Synthetic population of 10,000 parent-offspring trios
2. Logistic regression model to generate parent-offspring trait

# Local DTT Evaluation

## **Experimental setting**

1. Synthetic population of 10,000 parent-offspring trios
2. Logistic regression model to generate parent-offspring trait
3. FDR nominal level  $\alpha = 0.2$

# Local DTT Evaluation

## Experimental setting

1. Synthetic population of 10,000 parent-offspring trios
2. Logistic regression model to generate parent-offspring trait
3. FDR nominal level  $\alpha = 0.2$
4. 591, 513 SNPs on chromosomes 1 – 22, split into 532 pre-determined groups of size  $\sim 5$  Mb

# Local DTT Evaluation

## Experimental setting

1. Synthetic population of 10,000 parent-offspring trios
2. Logistic regression model to generate parent-offspring trait
3. FDR nominal level  $\alpha = 0.2$
4. 591, 513 SNPs on chromosomes 1 – 22, split into 532 pre-determined groups of size  $\sim 5$  Mb
5. DTT on each group and accumulation test to produce the set of discoveries

# Local DTT Evaluation

## Experimental setting

1. Synthetic population of 10,000 parent-offspring trios
2. Logistic regression model to generate parent-offspring trait
3. FDR nominal level  $\alpha = 0.2$
4. 591, 513 SNPs on chromosomes 1 – 22, split into 532 pre-determined groups of size  $\sim 5$  Mb
5. DTT on each group and accumulation test to produce the set of discoveries
6. Comparison with TDT / TDT-Screen algorithms applying statistical corrections to  $p$ -values

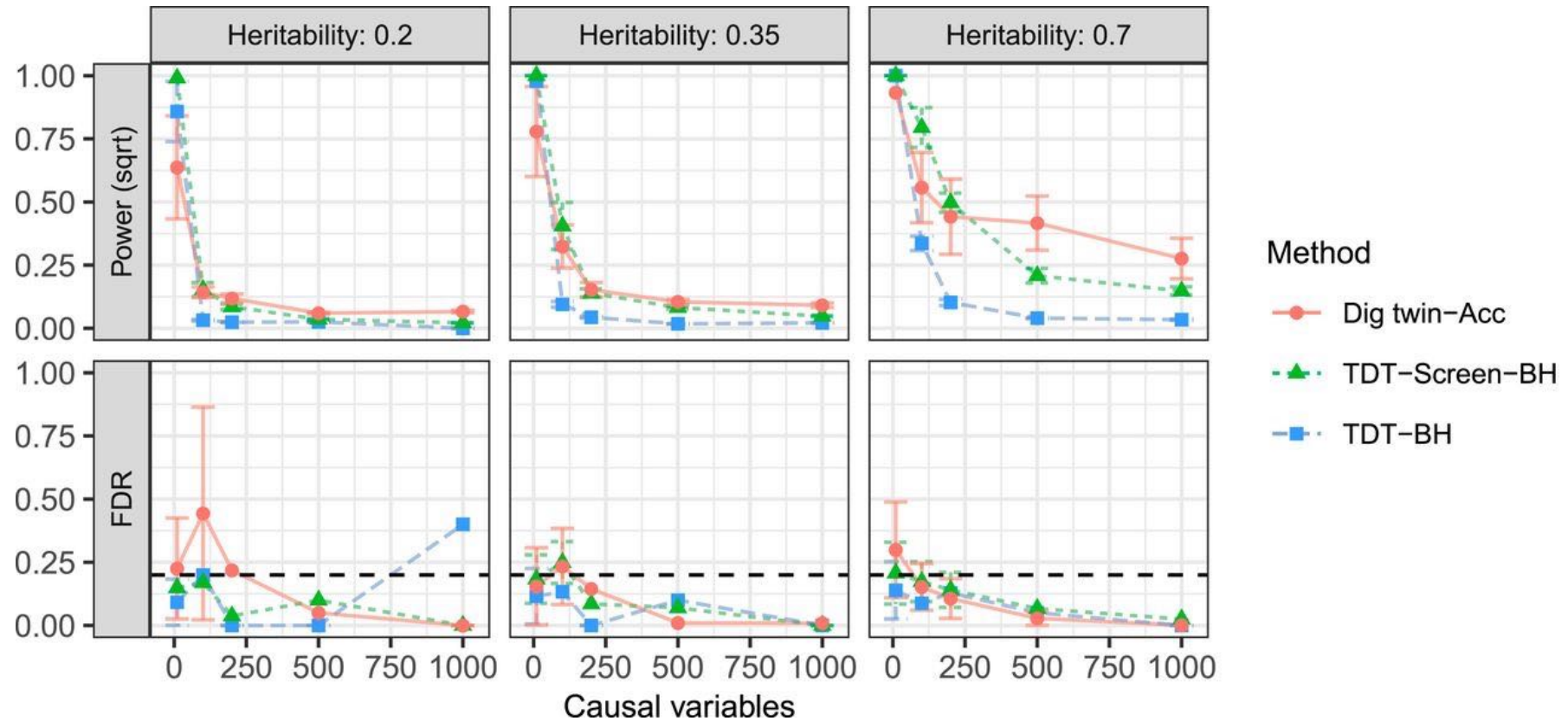
# Local DTT Evaluation

## Experimental setting

1. Synthetic population of 10,000 parent-offspring trios
2. Logistic regression model to generate parent-offspring trait
3. FDR nominal level  $\alpha = 0.2$
4. 591, 513 SNPs on chromosomes 1 – 22, split into 532 pre-determined groups of size  $\sim 5$  Mb
5. DTT on each group and accumulation test to produce the set of discoveries
6. Comparison with TDT / TDT-Screen algorithms applying statistical corrections to  $p$ -values
7. Benchmark algorithms do not have formal guarantees for localization due to full-chromosome null test



# Local DTT Evaluation



TDT / TDT-Screen lead to spurious discoveries because they cannot give reliable information about SNPs locality due to intra-chromosome SNPs dependency (linkage disequilibrium)

# Case Study: Autism Spectrum Disorder (ASD)

## Experimental setting

1. Local Digital Twin Test applied to a dataset of 2,565 parent-child trios to study whether the intergenic variant *rs910805* on chromosome 20 can be the cause of ASD
2. They applied Local DTT to groups centered around SNP *rs910805* of size ranging from 1Mb to full-chromosome
3. Significance level  $\alpha = 0.05$

# Case Study: Autism Spectrum Disorder (ASD)

<b>Resolution</b>	<b>1 Mb</b>	<b>2 Mb</b>	<b>3 Mb</b>	<b>4 Mb</b>	<b>5 Mb</b>	<b>full-chromosome</b>
<i>p</i> -value	0.237	0.146	0.100	0.0168	0.0244	0.011

Results show that  $H_0$  cannot be rejected at finer resolutions but is rejected for larger groups

# Case Study: Autism Spectrum Disorder (ASD)

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

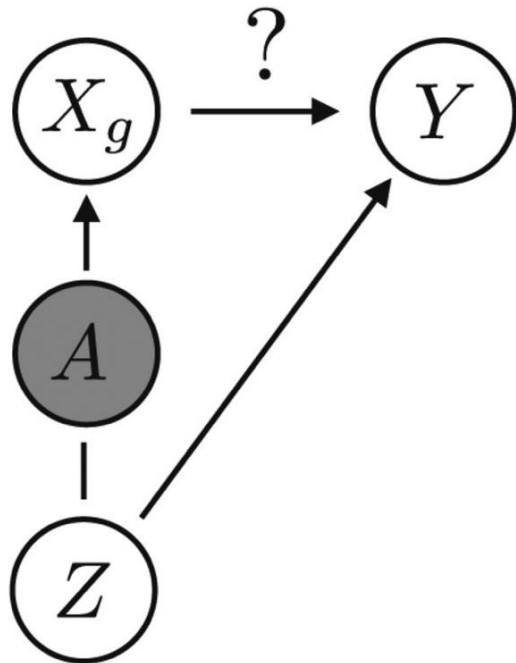
Results show that  $H_0$  cannot be rejected at finer resolutions but is rejected for larger groups



Observed association between the vicinity of SNP *rs910805* and ASD is not due to external confounders

# Final Remarks

DTT aims to establish if a genomic region contains causal SNPs



## Findings

1. the haplotypes  $A$  block the external confounders
2. given  $A$  we exactly know the distribution of  $X$
3. given such causal model if  $A$  satisfy the backdoor criterion w.r.t.  $X$  and  $Y$ ,  
 $H_0: X_g \perp\!\!\!\perp Y \mid A$  only detects causality

## Limits

1. computational phased haplotypes used by DTT can be subject to phased errors
2. in some case it might be harder to collect parent-offspring data than unrelated individuals



**Thank you!**

*Daniele Maria Papetti, Alessandro Tundo, Matteo Vaghi*