frontiers in Genetics

METHODS
published: 06 October 2015
doi: 10.3389/fgene.2015.00291

Simultaneous inference of haplotypes and alleles at a causal gene

Fabrice Larribe[1]*, *Mathieu J. Dupont*[2] and *Gabrielle Boucher*[3]

[1] Département de Mathématiques, Université du Québec à Montréal, Montréal, QC, Canada, [2] Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, QC, Canada, [3] Montreal Heart Institute, Montréal, QC, Canada

We present a methodology which jointly infers haplotypes and the causal alleles at a gene influencing a given trait. Often in human genetic studies, the available data consists of genotypes (series of genetic markers along the chromosomes) and a phenotype. However, for many genetic analyses, one needs haplotypes instead of genotypes. Our methodology is not only able to estimate haplotypes conditionally on the disease status, but is also able to infer the alleles at the unknown disease locus. Some applications of our methodology are in genetic mapping and in genetic counseling.

Keywords: EM algorithm, linkage disequilibrium, causal allele inference, haplotype estimation

OPEN ACCESS

Edited by:
Mariza De Andrade, Mayo Clinic, USA

Reviewed by:
Bjarni V. Halldorsson, Reykjavik University, Iceland
Tian-Qing Zheng, Chinese Academy of Agricultural Sciences, China

***Correspondence:**
Fabrice Larribe, Département de Mathématiques, Université du Québec à Montréal, Case Postale 8888, Succursale Centre-Ville, Montréal, QC H3C 3P8, Canada
larribe.fabrice@uqam.ca

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 February 2015*
Accepted: *02 September 2015*
Published: *06 October 2015*

Citation:
Larribe F, Dupont MJ and Boucher G (2015) Simultaneous inference of haplotypes and alleles at a causal gene. Front. Genet. 6:291. doi: 10.3389/fgene.2015.00291

1. Introduction

In human genetic studies, the unobserved raw data consists of two DNA sequences for each individual (see **Figure 1i**). As we usually observe few variations along the sequences for different people, we consider only sites which are different between individuals, the genetic markers. In this work we will consider bi-allelic markers, i.e., markers having two possible variations (alleles); the data can then be summarized as a binary couple. A series of genetic markers along the sequence is a haplotype; however, we observe genotypes, not haplotypes. In short, genotypes contain the same information as haplotypes, with the exception that they do not provide the phase information, i.e., we do not know which allele is located on which chromosome (see **Figure 1**, for an illustration). Moreover, we assume here that the phenotype (indicated by case/control) is caused or influenced by a binary (present or absent) Trait Influencing Mutation (TIM). The position of this TIM is always unknown; obviously, inferring the position of such a TIM is the precise goal of genetic mapping: to infer the location of such a TIM. In **Figure 1**, the TIM is illustrated as a DNA sequence of length six, however, in general TIMs can be sequences of any length, a single site for example.

For many genetic analyses however, the haplotype data is required, and in some cases even this information may not suffice. The unknown allelic state at the TIM is also required and this is an issue addressed by our work. Indeed, many methodologies use genealogies to infer parameters of populations (such as recombination rate and mutation rate) and these genealogies must be built using haplotypes, not genotypes, since the haplotypes contain the additional information about which genetic material was transmitted from one ancestor to a child. All such methodologies have a way to deal with this problem: some include an estimate of the haplotypes, like the Margarita program (Minichiello and Durbin, 2006), and others defer the issue to external methodologies, like TreeLD (Zöllner and Pritchard, 2005). We introduced an approach for fine genetic mapping using the coalescent with recombination (Larribe et al., 2002), which has the particularity to be the only methodology using genealogies built conditionally on the alleles at the TIM. Of course, as

the location of the TIM is unknown, the alleles at the TIM are unknown as well, and hence must be inferred from the data. Finally, the estimates of the alleles at the TIM opens new avenues to evaluate risks associated with the disease: under a standard genetic model, the risk of the disease is a function of the alleles at the TIM.

2. Existing Methodologies

To our knowledge, none of the current methodologies is able to jointly estimate haplotypes and the alleles at a causal gene. To infer haplotypes from genotypes, laboratory or computational methods can be used (Browning and Browning, 2011). The first statistical method to estimate haplotypes was the parsimony principle proposed by Clark (1990): assuming that the recombination rate is low, haplotypes are inferred by supposing that the number of distinct haplotypes is small. This method is limited to a few markers and works for short sequences. Since then, different methods have used the EM algorithm, for example Excoffier and Slatkin (1995) or Qin et al. (2002). If we wish to estimate the distribution of the haplotypes, we are in an incomplete data setting, and the EM algorithm is a natural solution. Indeed, this algorithm has often been used in this context; for example, one of the earliest methods to use haplotypes frequencies estimated by the EM algorithm is Fallin et al. (2001). One of the more advanced methods to estimate haplotypes is Phase (Stephens et al., 2001; Stephens and Donnelly, 2003); this bayesian method is based on Gibbs sampling and uses coalescent theory to derive the prior distribution of haplotypes. Other methods use information from reference haplotypes to improve phase estimation and infer missing genotypes. A comparison of the efficiency of different methods has been realized by Marchini et al. (2006) and Browning and Browning (2011), provide a recent review of methods that infer haplotypes.

Most of the aforementioned methodologies do not take into account the phenotype, nor the genetic model. In our context of a case/control study, one would not want to ignore the phenotype information. Unlike our method, none of the existing methodologies proposes to estimate the alleles at the TIM.

Finally, note that recent methodologies for estimating haplotypes use human sequence data, but some parts of the human genome are still difficult to sequence, which can limit the use of these methods; it is important to note that our method can be used on human sequences as well as animal or plant sequences, could also be extended to non diploid organisms.

3. The EM Algorithm

Note that the EM algorithm presented here, by taking into account the phenotype, is in a way an extension of the work of Excoffier and Slatkin (1995), and is also related to the work presented in chapter 5 of Foulkes (2009).

3.1. Complete Likelihood and M Step

Assume a large population of diploid individuals in Hardy-Weinberg equilibrium, where we can observe a dichotomous trait that depends on at least one Trait Influencing Mutation (TIM). As the phenotype ϕ depends on the TIM through a genetic model, it is actually possible for the trait to be dependent on several TIMs, but we consider only one TIM at a time. Let V_0 denote the distribution of haplotypes among non carriers of the TIM, and V_1 denote the distribution of haplotypes among carriers. Alternatively, carrier haplotypes will be called mutant haplotypes, and non carrier haplotypes, primitive haplotypes. For a given type of haplotype h ($h = 1, \ldots, H$), $V_0(h)$ is the proportion of haplotypes of type h among non carrier haplotypes, and similarly $V_1(h)$ is the proportion of haplotypes of type h among carrier haplotypes. A genetic model $F = (f_0, f_1, f_2)$ associated with the TIM we are studying is such that f_i is the probability for an individual to express the trait given that it bears $i = 0, 1$, or 2 copies of the causal mutation. Let T be the status of an individual at the TIM, such that $T \in \{(0,0), (1,1), (0,1), (1,0)\}$ represents a non carrier, an homozygote carrier or an heterozygote carrier with the TIM inherited from the father or the mother, respectively. Finally, let p be the frequency of carrier haplotypes in the population, and f the frequency of the trait we are working on.

FIGURE 1 | Illustration of a (very short) piece of chromosome (52 base pairs) for one individual; from top to bottom: (i) the unobserved raw data are the 2 DNA sequences; genetic markers in red, TIM in blue; (ii) haplotypes obtained by considering only genetic variations in DNA, and the unobserved TIM; (iii) finally, the observed data: a set of genotypes and a phenotype, in this case a status of a diseased individual.

Since the population is in Hardy-Weinberg equilibrium, we can easily calculate specific probabilities of the form $P[\phi, T = (\delta_1, \delta_2)]$; the sample spaces for T and ϕ being of size four and two respectively, there are eight such probabilities for their combinations. For example, as p^2 is the probability for an individual to be a double carrier, and f_2 the probability for a double carrier to express the trait, the probability for any individual i in the population to be a double carrier and to express the trait is simply

$$P\left[\phi_i = 1, T = (1,1)\right] \quad = \quad p^2 f_2. \tag{1}$$

The other seven cases are treated similarly (see **Table 1**, for details).

Let h^0 be a non carrier haplotype of type h, h^1 a carrier haplotype of type h, and $d_i = (h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2})$ the diplotype of individual i ($i = 1, \ldots, n$), where $\delta_1, \delta_2 \in \{0, 1\}$. Let G be a sample of genotypes from the population, and Φ the associated set of phenotypes. As we need to estimate V_0 and V_1, we are in an incomplete data problem, where the complete data is the set of phenotypes Φ and the set of diplotypes D, including the alleles at the causal gene. If the alleles (δ_1, δ_2) at the causal gene were known, then the probability of the diplotype $d_i = (h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2})$ would only depend on the distributions V_0 and V_1, and we would then have:

$$P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \mid V_0, V_1\right] = P\left[T = (\delta_1, \delta_2)\right]\ V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2}).$$

Since the phenotype depends on the diplotype only through the causal gene, the joint probability of the diplotype $d_i = (h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2})$ and the phenotype is:

$$\begin{aligned} &P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0, V_1\right] \\ &\quad = P\left[\phi_i \mid \left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right)\right] P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \mid V_0, V_1\right] \\ &\quad = P\left[\phi_i \mid T = (\delta_1, \delta_2)\right] P\left[T = (\delta_1, \delta_2)\right]\ V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2}) \\ &\quad = P\left[\phi_i, T = (\delta_1, \delta_2)\right] V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2}). \end{aligned} \tag{2}$$

We have previously seen how to calculate probabilities of the form $P[\phi, T = (\delta_1, \delta_2)]$ (see for example Equation 1); hence it becomes easy to calculate the above probability for each of the eight combinations of δ_1, δ_2 and ϕ; for instance:

$$\begin{aligned} P\left[\left(h_i^1, h_j^1\right), \phi = 1 \mid V_0, V_1\right] &= f_2 \cdot p^2 \cdot V_1(h_i) \cdot V_1(h_j), \\ P\left[\left(h_i^0, h_j^1\right), \phi = 0 \mid V_0, V_1\right] &= (1 - f_1) \cdot p \cdot (1 - p) \cdot \\ &\qquad V_0(h_i) \cdot V_1(h_j). \end{aligned}$$

Because individuals are assumed independent, the likelihood of (V_0, V_1) on the complete data is:

$$\begin{aligned} L_c(V_0, V_1) &= P\left[D, \Phi \mid V_0, V_1\right] \\ &= \prod_{i=1}^{n} P\left[d_i, \phi_i \mid V_0, V_1\right] \\ &= \prod_{i=1}^{n} P\left[\left(h_{i_1}^{\delta_{i_1}}, h_{i_2}^{\delta_{i_2}}\right), \phi_i \mid V_0, V_1\right] \\ &= \prod_{i=1}^{n} P\left[\phi_i, T = \left(\delta_{i_1}, \delta_{i_2}\right)\right] V_{\delta_{i_1}}(h_{i_1})\ V_{\delta_{i_2}}(h_{i_2}). \end{aligned}$$

TABLE 1 | Distributions of alleles at the causal gene in the population.

	Case	Control	Total
$T = (0,0)$	$f_0(1-p)^2$	$(1-f_0)(1-p)^2$	$(1-p)^2$
$T = (0,1)$	$f_1 p(1-p)$	$(1-f_1)p(1-p)^2$	$p(1-p)^2$
$T = (1,0)$	$f_1 p(1-p)$	$(1-f_1)p(1-p)^2$	$p(1-p)^2$
$T = (1,1)$	$f_2 p^2$	$(1-f_2)p^2$	p^2
Total	f	$1-f$	1

Since the probabilities $P[\phi_i, T = (\delta_{i1}, \delta_{i2})]$ do not depend on the distributions V_0 and V_1 but only on the penetrance model F and on the frequency p of carrier haplotypes, by denoting $K(F, p)$ a function that depends only on F and p, we have:

$$\begin{aligned} L_c(V_0, V_1) &= K(F,p) \prod_{i=1}^{n} V_{\delta_{i_1}}(h_{i_1})\ V_{\delta_{i_2}}(h_{i_2}) \\ &= K(F,p) \prod_{i=h}^{H} V_0(h)^{m_{h^0}}\ V_1(h)^{m_{h^1}}, \end{aligned}$$

where the last expression is obtained by taking the product over the types of haplotypes instead of individuals, and m_{h^0} and m_{h^1} are, respectively, the numbers of non carrier and carrier sequences of type h in D. This likelihood belongs to an exponential family, where the sufficient statistics for V_0 and V_1 are the frequencies m_{h^0} and m_{h^1}. If diplotypes were known, we could obtain the theoretical frequencies from the empirical ones. The diplotypes are not observable, but could be estimated if V_0 and V_1 were known.

Denote the expectation of the sufficient statistics by:

$$m_{h^\delta}^{(k+1)} = E\left[m_{h^\delta} \mid V_0^{(k)}, V_1^{(k)}, G, \phi\right].$$

We then have to maximize the function:

$$\begin{aligned} &W\left(V_0, V_1 \mid V_0^{(k)}, V_1^{(k)}\right) \\ &\quad = \sum_h \left[m_{h^0}^{(k+1)} \log(V_0(h)) + m_{h^1}^{(k+1)} \log(V_1(h))\right] \end{aligned}$$

with the constraints $\sum_h V_0(h) = 1$ and $\sum_h V_1(h) = 1$. By incorporating a Lagrange multiplier for each of the two constraints, we then have to optimize the linear expression :

$$\begin{aligned} &W_L\left(V_0, V_1 \mid V_0^{(k)}, V_1^{(k)}\right) \\ &\quad = \sum_h \left[m_{h^0}^{(k+1)} \log(V_0(h)) + m_{h^1}^{(k+1)} \log(V_1(h))\right] \\ &\quad + \quad \lambda_0\left(1 - \sum\nolimits_h V_0(h)\right) + \quad \lambda_1\left(1 - \sum\nolimits_h V_1(h)\right), \end{aligned}$$

i.e., we obtain maximum likelihood estimates from the complete data. It can be shown that W_L is a maximum if

$$V_0(h) = \frac{m_{h^0}^{(k+1)}}{\lambda_0}, \qquad V_1(h) = \frac{m_{h^1}^{(k+1)}}{\lambda_1}.$$

Applying the constraints, the M step of the algorithm consists in evaluating

$$V_0(h)^{(k+1)} = \frac{m_{h^0}^{(k+1)}}{\sum_h m_{h^0}^{(k+1)}}, \qquad V_1(h)^{(k+1)} = \frac{m_{h^1}^{(k+1)}}{\sum_h m_{h^1}^{(k+1)}},$$

where $\sum_h m_{h^0}^{(k+1)}$ and $\sum_h m_{h^1}^{(k+1)}$ are, respectively, the expected numbers of non carrier and carrier sequences after iteration k.

3.2. Conditional Expectation and the E Step

For now we have seen how to evaluate $V_0(h)^{(k+1)}$ and $V_1(h)^{(k+1)}$ with the haplotypes' frequencies; further we have to evaluate the conditional expectations $m_{h^\delta}^{(k+1)} = E[m_{h^\delta} \mid V_0^{(k)}, V_1^{(k)}, G, \phi]$. We have seen in Equation (2) that

$$P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0, V_1\right] = P\left[\phi_i, T = (\delta_1, \delta_2)\right]\ V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2}),$$

which gives, by conditioning on the phenotype:

$$P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0, V_1\right] = P\left[T = (\delta_1, \delta_2) \mid \phi_i\right]\ P\left[\phi_i\right]\ V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2}). \quad (3)$$

Using $P[\phi_i = 1] = f$, $P[\phi_i = 0] = 1 - f$ and the probabilities found earlier (see **Table 1**), it is immediate to obtain the probabilities of T given the phenotype, for example:

$$P\left[T_i = (0,0) \mid \phi_i = 1\right] = \frac{f_0(1-p)^2}{f},$$
$$P\left[T_i = (0,1) \mid \phi_i = 0\right] = \frac{(1-f_1)p(1-p)^2}{1-f}. \quad (4)$$

The joint probability of the genotype g_i and the phenotype ϕ_i is obtained from Equation (3) by summing over all the possible diplotypes:

$$P\left[g_i, \phi_i \mid V_0, V_1\right] = \sum_{\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \in g} P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0, V_1\right] = P\left[\phi_i\right] \sum_{\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \in g} P\left[T = (\delta_1, \delta_2) \mid \phi_i\right] V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2}). \quad (5)$$

Then, the probability of a diplotype $(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2})$, given the phenotype ϕ_i and the genotype g_i, is, using Equations (3) and (5):

$$P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \mid g_i, \phi_i, V_0, V_1\right] = \frac{P\left[T = (\delta_1, \delta_2) \mid \phi_i\right] V_{\delta_1}(h_{i_1})\ V_{\delta_2}(h_{i_2})}{\sum_{\left(h_{i_1}^{\beta_1}, h_{i_2}^{\beta_2}\right) \in g} P\left[T = (\beta_1, \beta_2) \mid \phi_i\right]\ V_{\beta_1}(h_{i_1})\ V_{\beta_2}(h_{i_2})}. \quad (6)$$

We can see that the conditional probability depends only on the distributions V_0 and V_1, and the probabilities $P[T \mid \phi_i]$.

Let's now evaluate the conditional expectation $m_{h^\delta}^{(k+1)}$. Let $n_{g,\phi}$ be the number of individuals with genotype g and phenotype ϕ; among these individuals,

$$n_{g,\phi} \cdot P\left[\left(h^\delta, \cdot\right) \mid g, \phi, \mid V_0, V_1\right]$$

will receive the sequence h^δ from their mother, and

$$n_{g,\phi} \cdot P\left[\left(\cdot, h^\delta\right) \mid g, \phi, \mid V_0, V_1\right]$$

from their father. As usual, the conditional probability of having a given sequence as the maternal haplotype is obtained by summing on all the compatible paternal haplotypes (and vice versa). Recall that if there is no missing information on the genotypes, there is a unique sequence h_g compatible with h such that $(h, h_g) \in g$. In this case:

$$P\left[\left(h^\delta, \cdot\right) \mid g, \phi, V_0, V_1\right] = P\left[\left(h^\delta, h_g^0\right) \mid g, \phi, V_0, V_1\right] + P\left[\left(h^\delta, h_g^1\right) \mid g, \phi, V_0, V_1\right],$$
$$P\left[\left(\cdot, h^\delta\right) \mid g, \phi, V_0, V_1\right] = P\left[\left(h_g^0, h^\delta\right) \mid g, \phi, V_0, V_1\right] + P\left[\left(h_g^1, h^\delta\right) \mid g, \phi, V_0, V_1\right].$$

These two probabilities being equal by symmetry, the mean number of copies of h^δ carried by the $n_{g,\phi}$ individuals presenting this profile is

$$2 \cdot n_{g,\phi} \cdot P\left[\left(h_g, \cdot\right) \mid g, \phi, V_0, V_1\right].$$

We then obtain $m_{h_\delta}^{(k+1)}$ by summing over all the genotypes and phenotypes. The E step of the algorithm reduces to evaluating:

$$m_{h^\delta}^{(k+1)} = \sum_{(g,\phi) \in (G,\Phi)} 2 \cdot n_{g,\phi} \cdot P\left[\left(h_g, \cdot\right) \mid g, \phi, V_0, V_1\right]. \quad (7)$$

for each h^δ. Note that the method can be generalized to missing data, by considering every combination of haplotypes compatible with the observed genotypes.

3.3. Non Random Sampling

We have assumed until now that the sample was obtained by simple random sampling from the population, but usually this is not the case in genetics, since most samples are obtained using a case/control design. Let n_1 be the number of cases in the sample of size n, and let $\omega = n_1/n$ be the proportion of cases (if the sample were a simple random sample, then ω is expected

TABLE 2 | Distributions of alleles at the causal gene in the population in a sample with a fixed proportion of cases.

	Case	Control	Total
$T = (0,0)$	$f_0(1-p)^2 \cdot \frac{\omega}{f}$	$(1-f_0)(1-p)^2 \cdot \frac{(1-\omega)}{(1-f)}$	q_{00}
$T = (0,1)$	$f_1 p(1-p) \cdot \frac{\omega}{f}$	$(1-f_1)p(1-p)^2 \cdot \frac{(1-\omega)}{(1-f)}$	q_{01}
$T = (1,0)$	$f_1 p(1-p) \cdot \frac{\omega}{f}$	$(1-f_1)p(1-p)^2 \cdot \frac{(1-\omega)}{(1-f)}$	q_{10}
$T = (1,1)$	$f_2 p^2 \cdot \frac{\omega}{f}$	$(1-f_2)p^2 \cdot \frac{(1-\omega)}{(1-f)}$	q_{11}
Total	ω	$1-\omega$	1

to be $P[\phi] = f$). In the case/control setting, many methods which estimate haplotypes, including those which use the EM algorithm, are biased since the case/control mode of sampling modifies the distributions of alleles and haplotypes.

In this section we show that the algorithm described in Section 2.2 is robust to this case/control sampling. The proportions given in **Table 1**, as well as the penetrance model, are not affected by the sampling. Let's see in a first step the behavior of some probabilities under this stratified sampling design. The probabilities of T conditional on the phenotype do not change, i.e., $P[T \mid \phi, n_1] = P[T \mid \phi]$, and the probabilities defined previously are still valid. Expected distributions are then easily obtained (see **Table 2**).

Let's now review the steps of the algorithm. The likelihood on the complete data for such a stratified sample, conditional on the number of cases, is:

$$\begin{aligned} L_c(V_0, V_1) &= P[D, \Phi \mid V_0, V_1, n_1] \\ &= \frac{P[D, \Phi, V_0, V_1, n_1]}{P[n_1 \mid V_0, V_1, n_1]\ P[V_0, V_1]} \\ &= \frac{P[D, \Phi, n_1 \mid V_0, V_1]}{P[n_1 \mid V_0, V_1, n_1]\ P[V_0, V_1]}. \end{aligned}$$

Since knowledge of the phenotypes Φ carries knowledge of n_1, n_1 can be removed from the numerator. Moreover, the probability of obtaining n_1 cases from a simple random sample does not depend on the distributions V_0 and V_1. After removing terms which do not depend on V_0 and V_1, the likelihood for this data is the same as before:

$$\begin{aligned} L_c(V_0, V_1) &= \frac{P[D, \Phi, n_1 \mid V_0, V_1]}{P[n_1]} \\ &= \frac{P[D, \Phi \mid V_0, V_1]}{\binom{n}{n_1} f^{n^1}(1-f)^{n-n_1}} \\ &= K\left(F, p, n_1\right) \prod_h V_0(h)^{m_{h^0}}\ V_1(h)^{m_{h^1}}, \end{aligned}$$

which shows the likelihood remains the same for case/control sampling, and hence the M step of the algorithm remains unchanged.

Recall that the E step depended on diplotypes' probabilities, conditional on the genotype and the phenotype. We prove that these probabilities are not modified by the type of sampling. Let's begin by calculating the joint probability of a diplotype and a phenotype, conditional on ω, the proportion of cases; this probability is obtained by adding a condition on the proportion of cases in Equation (3):

$$\begin{aligned} &P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0, V_1, \omega\right] \\ &\quad = P[\phi_i \mid n_1]\, P[T = (\delta_1, \delta_2) \mid \phi_i, \omega] \\ &\quad \times P\left[\left(h_{i_1}, h_{i_2}\right) \mid T = (\delta_1, \delta_2), \omega\right]. \end{aligned}$$

Once the status at the causal gene is determined, the diplotype probability depends only on the distributions V_0 and V_1. Moreover, if the phenotype is known, T does not depend on ω; only the probability of the phenotype is modified:

$$\begin{aligned} &P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0, V_1, \omega\right] \\ &\quad = P[\phi_i \mid \omega]\, P[T = (\delta_1, \delta_2) \mid \phi_i]\, V_{\delta_1}(h_{i_1})\, V_{\delta_2}(h_{i_2}). \end{aligned}$$

Following the derivation for a simple random sample, the term $P[\phi_i \mid \omega]$ cancels out in the conditional probability formula, and we get the same result as before. Because these probabilities are not affected by the sampling design, the E step of the algorithm, described in Equation (8) remains the same. We have shown that this EM algorithm can be applied to case/control samples.

3.4. Overview of the Algorithm

Assume the penetrance model $F = (f_0, f_1, f_2)$ and the frequency of the causal mutation are known. The steps of our algorithm are:

1. Compute the probabilities $P[T = (i, j) \mid \phi = \delta]$ (see Equation 4), for $(i, j, \delta) \in \{0, 1\}$;
2. Consider an initial $V_0^{(0)}$ and $V_1^{(0)}$ probability distribution (in absence of a priori information on the frequency of haplotypes, we use a uniform distribution);
3. E Step:
   (a) For each genotype g and phenotype ϕ (see Equation 3):
      (1) Evaluate, for all $[h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}] \in g$:

$$\begin{aligned} &P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0^{(k)}, V_1^{(k)}\right] \\ &\quad \propto V_{\delta_1}^{(k)}(h_{i_1})\, V_{\delta_2}^{(k)}(h_{i_2})\, P[T = \delta_1\delta_2 \mid \phi] \end{aligned}$$

      (2) Sum these probabilities to obtain:

$$\begin{aligned} &P\left[g, \phi_i \mid V_0^{(k)}, V_1^{(k)}\right] \\ &\quad = \sum_{\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \in g} P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi \mid V_0^{(k)}, V_1^{(k)}\right]. \end{aligned}$$

      (3) The conditional probability can then be computed as:

$$\begin{aligned} &P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right) \mid g, \phi, V_0^{(k)}, V_1^{(k)}\right] \\ &\quad = \frac{P\left[\left(h_{i_1}^{\delta_1}, h_{i_2}^{\delta_2}\right), \phi_i \mid V_0^{(k)}, V_1^{(k)}\right]}{P\left[g, \phi \mid V_0^{(k)}, V_1^{(k)}\right]}. \end{aligned}$$

(b) Compute, for each sequence h^δ (see Equation 8):

$$m_{h^\delta}^{(k+1)} = \sum_{(g,\phi)\in(G,\Phi)} 2 \cdot n_{g,\phi} \cdot P\left[\left(h_g, \cdot\right) \mid g, \phi, V_0^{(k)}, V_1^{(k)}\right]. \quad (8)$$

4. M Step: update the $V_\cdot$ distributions, by evaluating, for all h (see Equation 3.1):

$$V_0(h)^{(k+1)} = \frac{m_{h^0}^{(k+1)}}{\sum_h m_{h^0}^{(k+1)}}, \qquad V_1(h)^{(k+1)} = \frac{m_{h^1}^{(k+1)}}{\sum_h m_{h^1}^{(k+1)}}.$$

5. Convergence test. Convergence is reached when

$$\max_{h,\delta} \left| V_\delta(h)^{(k+1)} - V_\delta(h)^{(k)} \right| < \epsilon.$$

One convergence is reached, let $\hat{V}_0 = V_0^{(k+1)}$ and $\hat{V}_1 = V_1^{(k+1)}$. Otherwise, go back to step 3.

We have assumed that the proportion of carrier haplotypes is known, which is of course not realistic in practice. Note however, by assuming that the penetrance model F is known, and that the frequency of the disease f is known as well, we can obtain p. Since

$$f = f_0(1-p)^2 + 2f_1p(1-p) + f_2p^2,$$

if $f_0 + f_2 - 2f_1 = 0$, then $p = (f - f_0)/2(f_1 - f_0)$. In general, however, we have:

$$p = \frac{f_0 - f_1 \pm \sqrt{f_1^2 - f_0f_1 + f(f_0 - 2f_1 + f2)}}{f_0 - 2f1 + f_2},$$

and there exists a solution in $[0, 1]$ which satisfies the penetrance model. If $0 \le f_0 \le f_1 \le f_2$, then the solution is unique. The methodology has been implemented in C++, and is available from the corresponding author.

For the proposed illustration, we have used `ms` (Hudson, 2002) to sample 10,000 chromosomes of approximate length 250 kb (simulated using $\rho = 100$), and randomly assigned pairs of chromosomes to form diploid individuals. One of the markers is chosen randomly such that its minimum allele frequency is approximately 0.10, and this marker will become the TIM. For each individual, a phenotype is then simulated using the two alleles at the TIM according to the genetic model $F = (0.05, 0.10, 0.80)$. Haplotypes are then mixed in order to obtain genotypes, and information on haplotypes is discarded. We sample 100 individuals and sequences of length 8 markers, so 2^8 haplotypes are possible in theory, but only 24 of them are compatible with the observed genotypes. **Figure 2** shows the estimates of vectors V_0 and V_1 for each of the 24 possible types of haplotypes in the sample. By comparing individual values, $V_1(1)$ and $\hat{V}_1(1)$ for example, we see that the estimates of the frequencies are very good. Note that the estimates of $V_0(\cdot)$ seem to be slightly better than those of $V_1(\cdot)$; this is due to the fact that we have more information on control haplotypes than on case haplotypes, because phenocopy causes many case haplotypes to be non carriers.

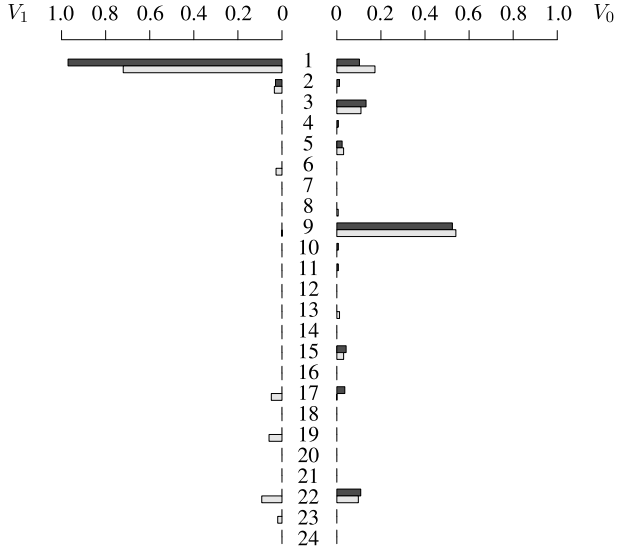

FIGURE 2 | Distribution of V_1 (left) and V_0 (right). Estimated distributions in light gray, and exact distributions in black. Each of the 24 horizontal bars represents a frequency of a particular haplotype type which is compatible with the data: of the 2^8 possible haplotypes of length 8 markers, only 24 were compatible with the observed genotypes.

4. Estimating the Causal Alleles

The methodology presented in this paper is the first to permit one to jointly estimate the haplotypes at genotyped markers and the (non observed) alleles at the TIM. Estimating the causal alleles could be very useful in genetic counseling for example, where the patient's risk and treatment could be adjusted if the alleles at the disease genes were known. In the sequel, we assume that haplotypes are known and we assess the capacity of our method to infer the causal alleles. As explained in Dupont (2012) or Boucher (2009), when the haplotypes are known, Equation (6) reduces to:

$$P\left[T = (\delta_1, \delta_2) \mid d = \left(h_1, h_2\right), \phi, V_0, V_1\right] = \frac{P\left[T = (\delta_1, \delta_2) \mid \phi\right] \; V_{\delta_1}(h_1)\, V_{\delta_2}(h_2)}{\sum_T P\left[T = (\delta_1, \delta_2) \mid \phi\right] \; V_{\delta_1}(h_1)\, V_{\delta_2}(h_2)}.$$

In the EM algorithm, the number of parameters increases as the number of genetic markers increases: since the markers are binary, if sequences of length d are used, there are 2^d possible haplotypes, leading to a maximum of $2^d - 1$ parameters to estimate. For this reason, and because huge numbers of genetic markers are available today, the method is illustrated here using a moving windows strategy, i.e., we use windows made of d markers each, and the total number of markers is L ($d < L$). The first window consists of the set of markers $\{1, 2, \ldots, d\}$, the second consists of the set of markers $\{2, \ldots, d+1\}$, and so on.

Let n_{cas} and n_{con} be the numbers of case and control haplotypes, and n^0 and n^1 the numbers of non carrier and carrier haplotypes, respectively. Let n_{cas}^1 and n_{con}^1 be the numbers of case and control carrier haplotypes, and n_{cas}^0 and n_{con}^0 the numbers of case and control non carrier haplotypes, respectively. We then

have $n = n^0 + n^1 = n_{\text{cas}} + n_{\text{con}} = n^0_{\text{cas}} + n^0_{\text{con}} + n^1_{\text{cas}} + n^1_{\text{con}}$. Finally, let $n^\delta_c(0)$ and $n^\delta_c(1)$ be the numbers of haplotypes, where the c subscript denotes the case/control status ($c \in \{\text{cas}, \text{con}\}$) and δ superscript ($\delta \in \{0, 1\}$) denotes the true non carrier/carrier status of the individuals in the counts $n^\delta_c(0)$ and $n^\delta_c(1)$. For example, out of a total of $n^1_{\text{cas}} = 100$ carrier cases, if 75 are correctly estimated as being carriers ($n^1_{\text{cas}}(1) = 75$), then 25 are erroneously estimated as being non carriers ($n^1_{\text{cas}}(0) = 25$), because $n^1_{\text{cas}}(1) + n^1_{\text{cas}}(0) = n^1_{\text{cas}}$. We then define the partial success rates:

$$\pi^0_{\text{con}} = \frac{n^0_{\text{con}}(0)}{n^0_{\text{con}}}, \qquad \pi^1_{\text{con}} = \frac{n^1_{\text{con}}(1)}{n^1_{\text{con}}},$$

$$\pi^0_{\text{cas}} = \frac{n^0_{\text{cas}}(0)}{n^0_{\text{cas}}}, \qquad \pi^1_{\text{cas}} = \frac{n^1_{\text{cas}}(1)}{n^1_{\text{cas}}},$$

and semi-partial success rates:

$$\pi_{\text{con}} = \frac{n^0_{\text{con}}(0) + n^1_{\text{con}}(1)}{n_{\text{con}}}, \qquad \pi_{\text{cas}} = \frac{n^0_{\text{cas}}(0) + n^1_{\text{cas}}(1)}{n_{\text{cas}}},$$

$$\pi^0 = \frac{n^0_{\text{con}}(0) + n^0_{\text{cas}}(0)}{n^0}, \qquad \pi^1 = \frac{n^1_{\text{con}}(1) + n^1_{\text{cas}}(1)}{n^1};$$

finally, we have the global success rate:

$$\pi = \frac{n^0_{\text{con}}(0) + n^1_{\text{con}}(1) + n^0_{\text{cas}}(0) + n^1_{\text{cas}}(1)}{n^0_{\text{con}} + n^1_{\text{con}} + n^0_{\text{cas}} + n^1_{\text{cas}}} = \frac{n^0(0) + n^1(1)}{n}.$$

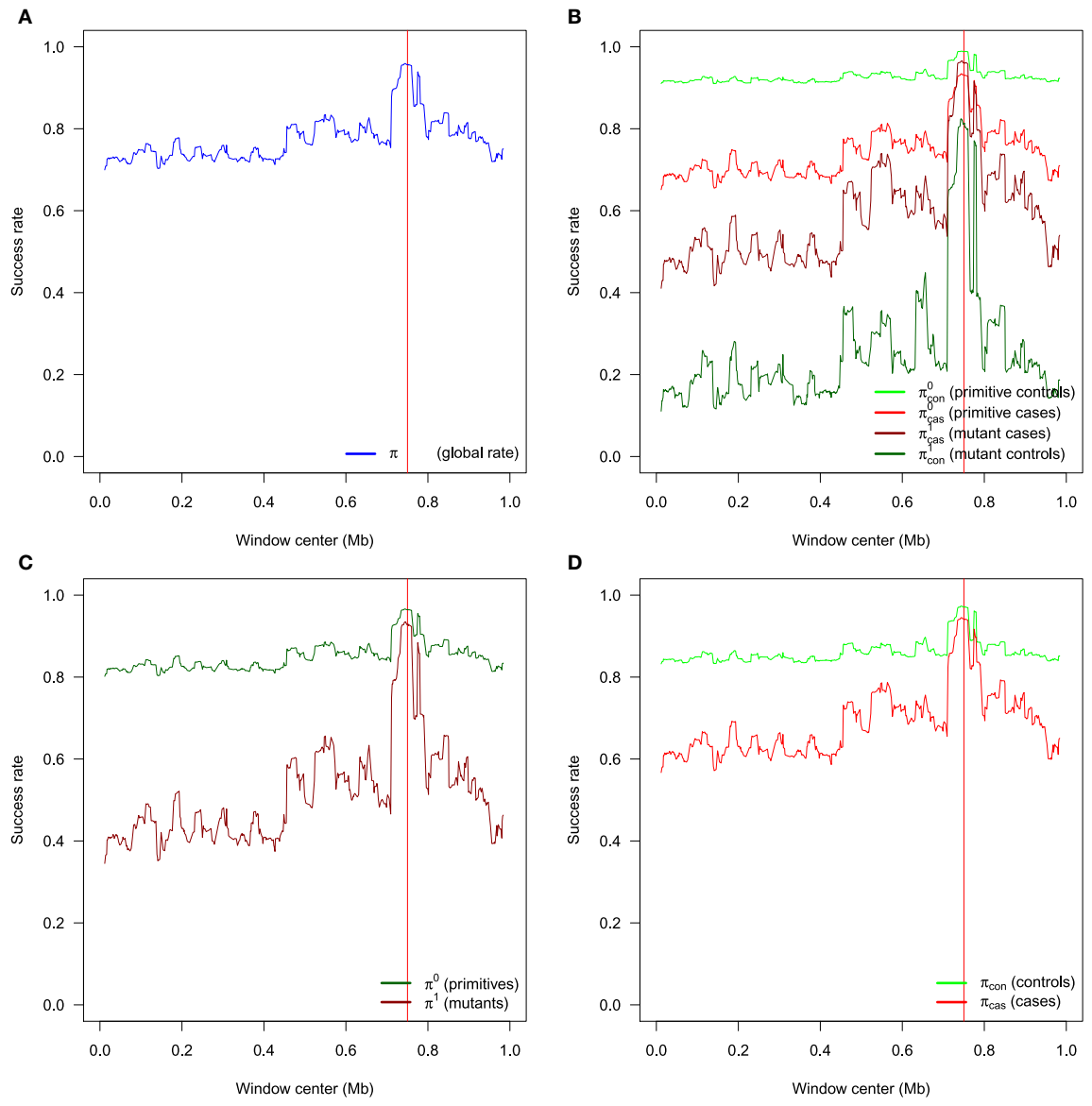

FIGURE 3 | Example of the different success rates for a particular sample of 400 cases and 400 controls obtained with the genetic model $F = (0.01, 0.1, 0.1)$. (A) Global success rate. **(B)** Partial success rates: primitive controls, primitive cases, mutant cases, mutant controls. **(C)** Semi-partial success rates: primitives and mutants. **(D)** Semi-partial success rates: controls and cases. Results are obtained with windows of width $d = 16$ SNPs incremented by one. There are a total of 500 SNPs.

All these rates have different meanings, and are useful depending on the question of interest. In particular, π^0 is the probability to estimate a non carrier if the individual is a non carrier, in other word the specificity, while π^1 is the probability to estimate a carrier if the individual is a carrier, in other word the sensitivity. The probability π is known as the accuracy. As with all classification rules, it is not informative to achieve high sensitivity without specificity and vice versa. Accuracy alone is not an ideal measure of success for low frequency TIM, as high accuracy could be achieved by simply setting $n_1 = 0$. **Figure 3** exhibits an example of the different success rates for a particular sample of 400 cases and 400 controls obtained with the genetic model $F = (0.01, 0.1, 0.1)$. Results are obtained with 500 SNPs using windows of width $d = 16$ SNPs incremented by one. The EM algorithm is run for each of the $L - d + 1$ windows along the sequence. The different success rates are estimated for each window and plotted at the center of the window. In each window along the sequence, the rate indicates the proportion of haplotypes for which the allele at the TIM is correctly inferred. In all figures, the real position of the TIM is indicated by a red vertical line. We can see in this example that there is variability along the sequences, and this is easier to estimate TIM alleles for primitive controls (light green) than mutant controls (dark green). All rates increase in the vinicity of the TIM; for instance, consider the global rate π (**Figure 3A**) which ranges from 0.70 to 0.96 at the position of the TIM. The mean global rate along the sequence is 0.77. The increased success rate near the TIM is due to more linkage disequilibrium around the TIM, and the difference in haplotypes between cases and controls is more informative.

The accuracy of our method depends on several factors. We have identified three of them: the genetic model, the sample size, and the windows width d. To identify the impact of each of these factors, we present further analyses where we vary the factors one at a time. In order to assess the effect on the estimation of the TIM's allele only, we assume here that the haplotypes are known. The strength of the genetic model will be measured using the risk

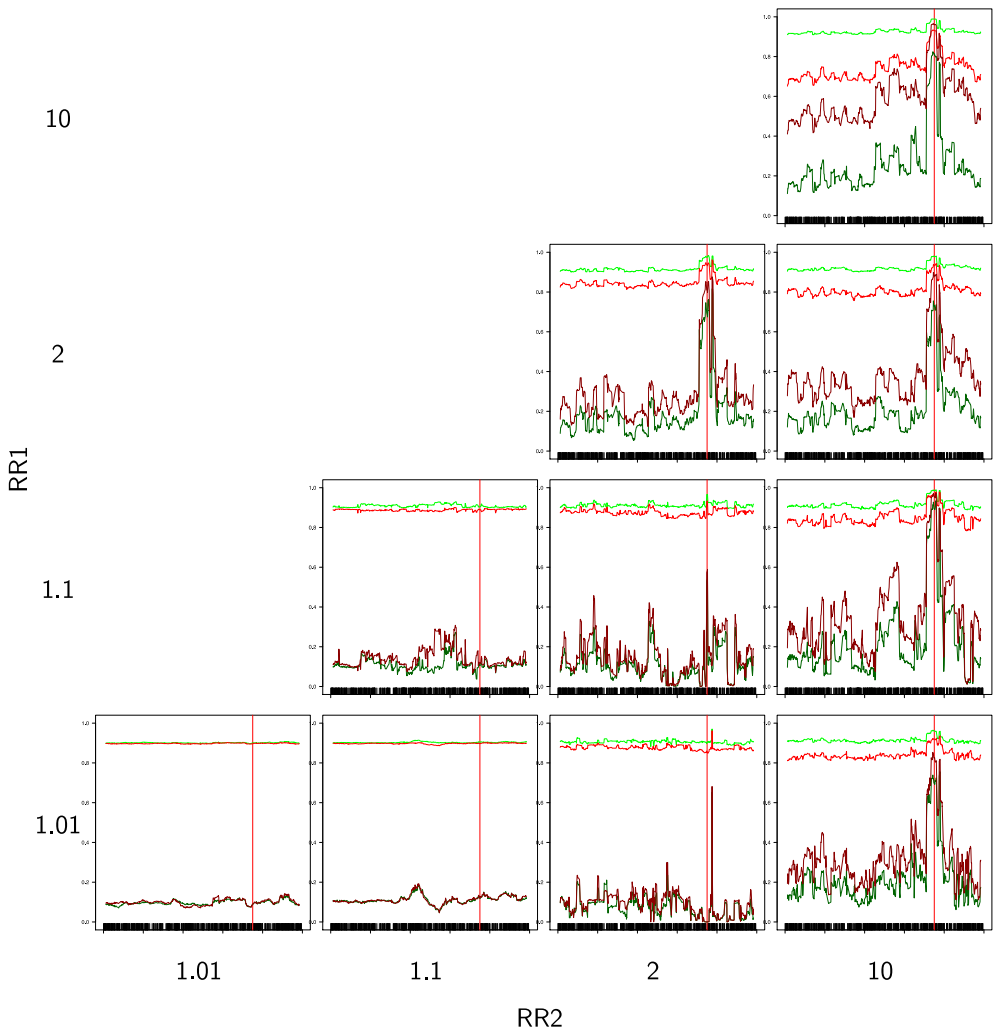

FIGURE 4 | Partial success rate depending on relative risks $RR_1 = RR_2 = 10$, for a sample size of 400/400 cases/controls, and windows of 16 SNPs. Scale on *Y* axis: [0, 1]. Light green: primitive controls (π^0_{con}), red: primitive cases (π^0_{cas}), dark red: mutant cases (π^1_{cas}), dark green: mutant controls (π^1_{con}).

ratio, RR, which is the ratio of the risk of one specified genotype compared to the genotype with no carrier allele. On a simulated population of 50,000 haplotypes and 10,000 SNPs generated by FastSimCoal (Excoffier et al., 2013), 40 of datasets are produced for various genetic models ($RR_1 = f_1/f_0 \in \{1.01, 1.1, 2, 10\}$, $RR_2 = f_2/f_0 \in \{1.01, 1.1, 2, 10\}$, for different sample sizes ($n_{con}/n_{cas} \in \{100/100, 200/200, 400/400, 800/800\}$) and for various window of widths $d \in \{2, 4, 8, 16\}$. Low relative risks RR_1 and RR_2 implies there is less information in the data to infer mutant haplotypes. To obtain an informative range of values for RR_1 and RR_2, we fixed f_0 at 0.01 and allowed f_1 and f_2 to take every value in the set $\{0.0101, 0.011, 0.02, 0.1\}$ such that $f_0 \leq f_1 \leq f_2$. These combinations lead to various genetic models, including recessive and dominant ones. Regarding the windows width, we expect that short windows contain less information about the data, however very large windows can cause many single haplotypes, making V^0 and V^1 difficult to estimate. Each sample originates from the same population, with a TIM

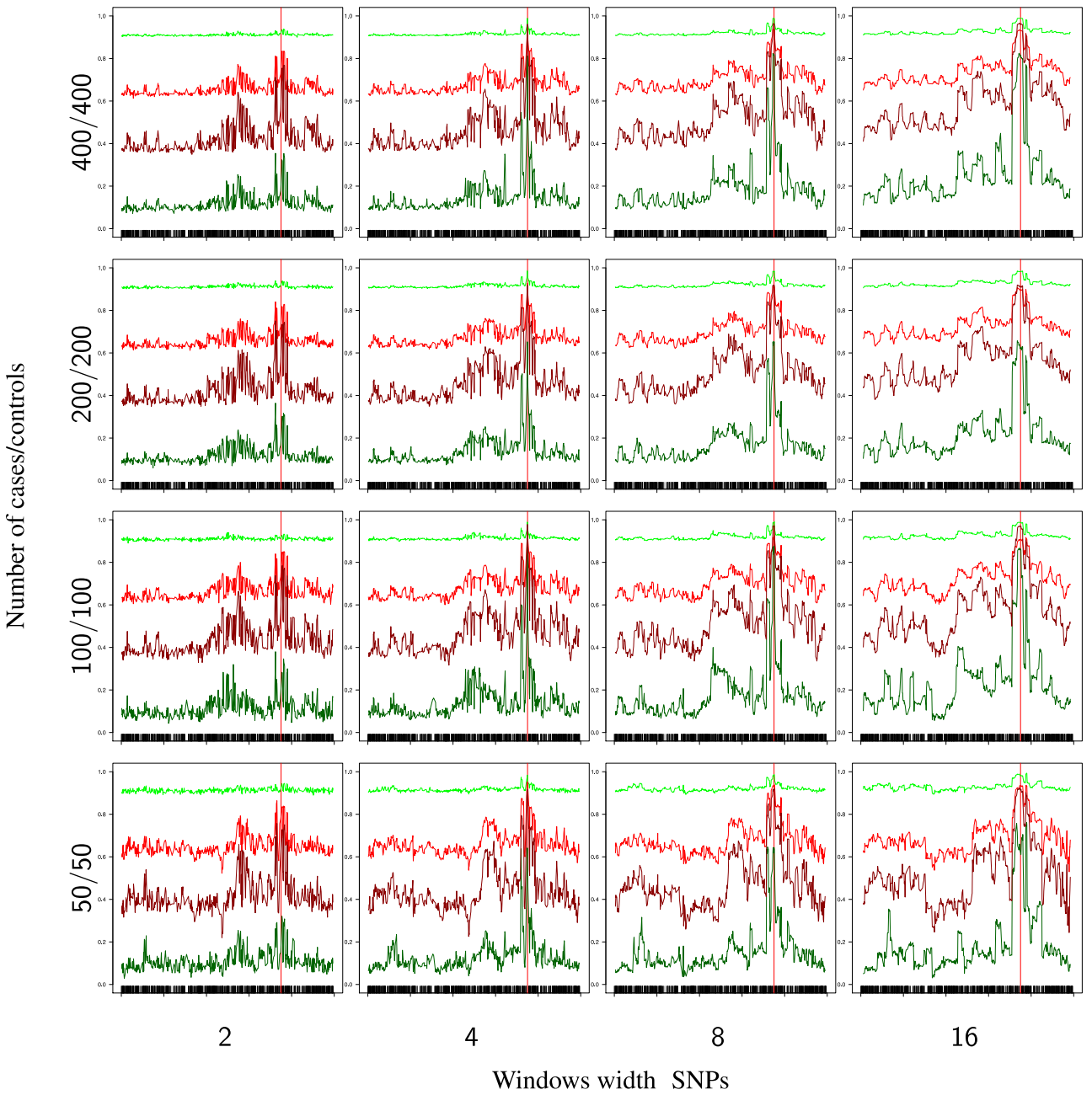

FIGURE 5 | Partial success rates depending on sample size and windows width for $RR_1 = RR_2 = 10$. Scale on the *Y* axis: [0, 1]. Light green: primitive controls (π^0_{con}), red: primitive cases (π^0_{cas}), dark red: mutant cases (π^1_{cas}), dark green: mutant controls (π^1_{con}).

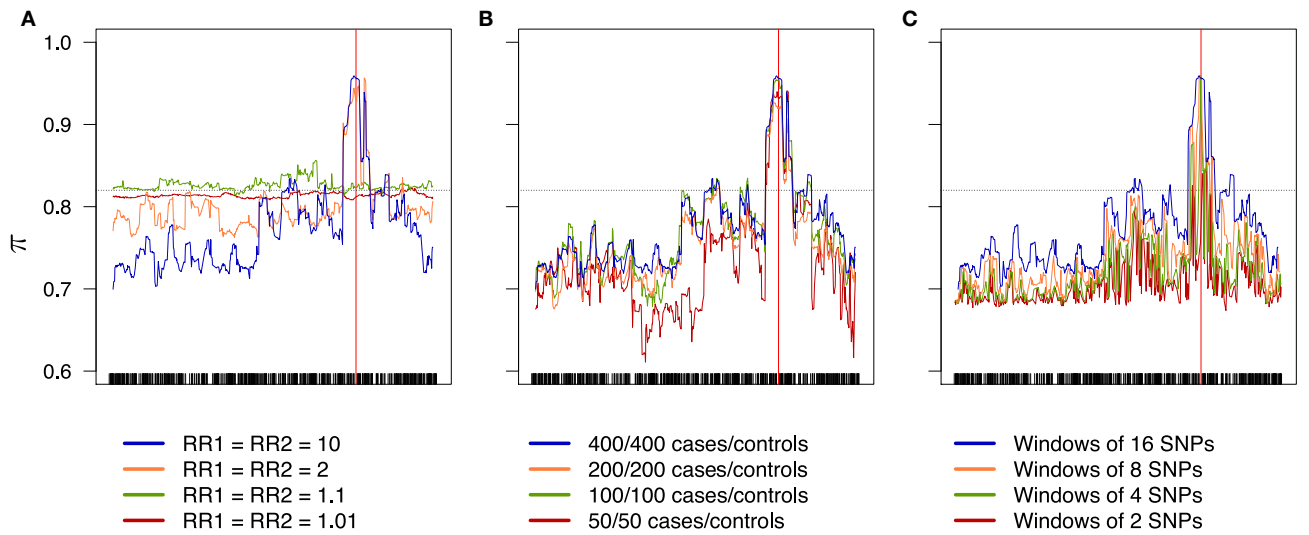

FIGURE 6 | Global success rates depending on combined relative risks (A), sample size (B), and windows width (C). Scale is the same for all three plots. The dotted line represents the random success rate (0.82). **(A)** 400/400 cases/controls, windows of 16 SNPs. **(B)** $RR_1 = RR_2 = 10$, windows of 16 SNPs. **(C)** $RR_1 = RR_2 = 10$, 400/400 cases/controls.

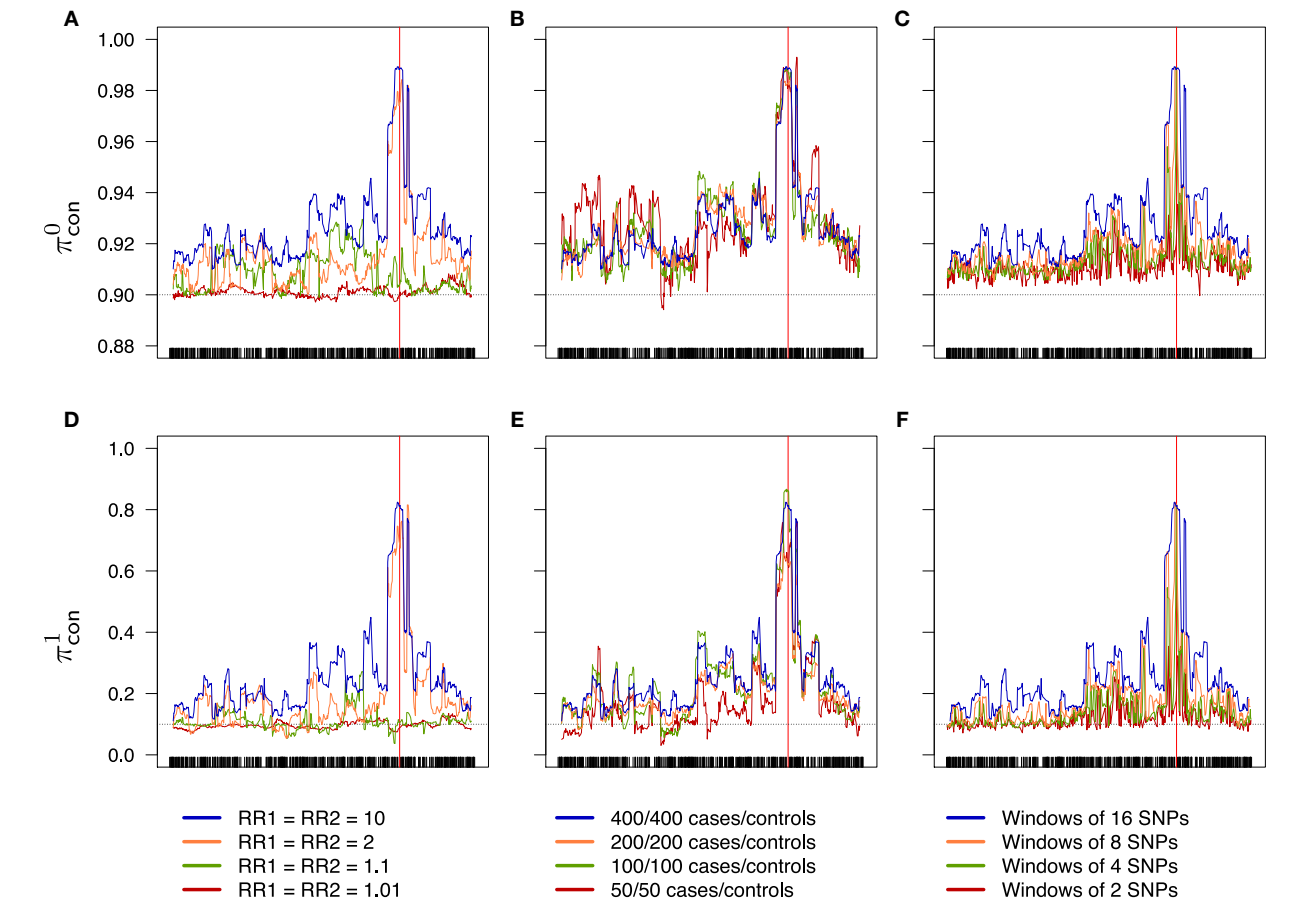

FIGURE 7 | Success rates of primitive controls (A–C) and mutant controls (D–F) depending on combined relative risks (A,D), sample size (B,E) and windows width (C,F). Scale is the same for a given line. Dotted lines represent the random success rate (**A–C**: 0.9; **D–F**: 0.1). **(A,D)** 400/400 cases/controls, windows of 16 SNPs. **(B,E)** $RR_1 = RR_2 = 10$, windows of 16 SNPs. **(C,F)** $RR_1 = RR_2 = 10$, 400/400 cases/controls.

frequency of $p = 0.1$. The same 500 selected SNPs are used for all analyses. The disease frequency in the population, is calculated as $f_0 \cdot (1-p)^2 + f_1 \cdot 2 \cdot p(1-p) + f_2 \cdot p^2$, and ranges in values from 0.01 to 0.0271. In a case control design with equal proportions of cases and controls, the frequency of the disease itself will have no effect on the proposed methodologies, but the genetic models will.

Figure 4 provides the partial success rates for different combinations of RR_1 and RR_2. When relative risks are low, we observe that the rates are constant along the sequences, with π^0_{con} and π^0_{cas} very high and π^1_{con} and π^1_{cas} very low (in fact, these rates are near $p = 0.1$ and $1 - p = 0.9$, which are the random success rates without any genetic effect). As soon as the relative risk increases, we observe an improvement in the estimation of the causal alleles, this improvement being very noticeable in region of the TIM.

The effect of the windows width (d) and the sample size are shown in **Figure 5**. There is a general improvement in the success rates (π^0_{con}, π^0_{cas}, π^1_{cas}, π^1_{con}) when these two factors increase, but, surprisingly, the effect is not very strong, perhaps due to the high relative risk assumed in these analyses.

An overview of the effect of the different factors on the global rate π (the accuracy) is shown in **Figure 6**; an increase in the relative risks clearly translates to an improvement of the global success rate. If the sample size is larger than 50/50 for cases/controls, increasing the sample size seems to have little effect, at least with the parameters considered in this example. Finally, the effect of the windows width is pretty clear: π increases all along the sequence when the width of the windows increases. Similar results are obtained for the partial rates π^1_{con}, π^0_{con} (see **Figure 7**) and for the partial rates π^1_{cas}, π^0_{cas} (see **Figure 8**).

We have compared this EM methodology to a simpler naive method, which consists of testing the association of each marker in the region with the phenotype, and to infer the alleles at the TIM to be the alleles at the marker having the strongest association with the phenotype. To illustrate this procedure, we have used 7503 heterozygote markers to test the association on the data in the case $RR_1 = RR_2 = 2$. As shown in **Figure 9A**, for each of the four markers showing the strongest association with the phenotype, we have inferred the TIM's alleles to be the alleles of this marker, and plotted the success rate π at the

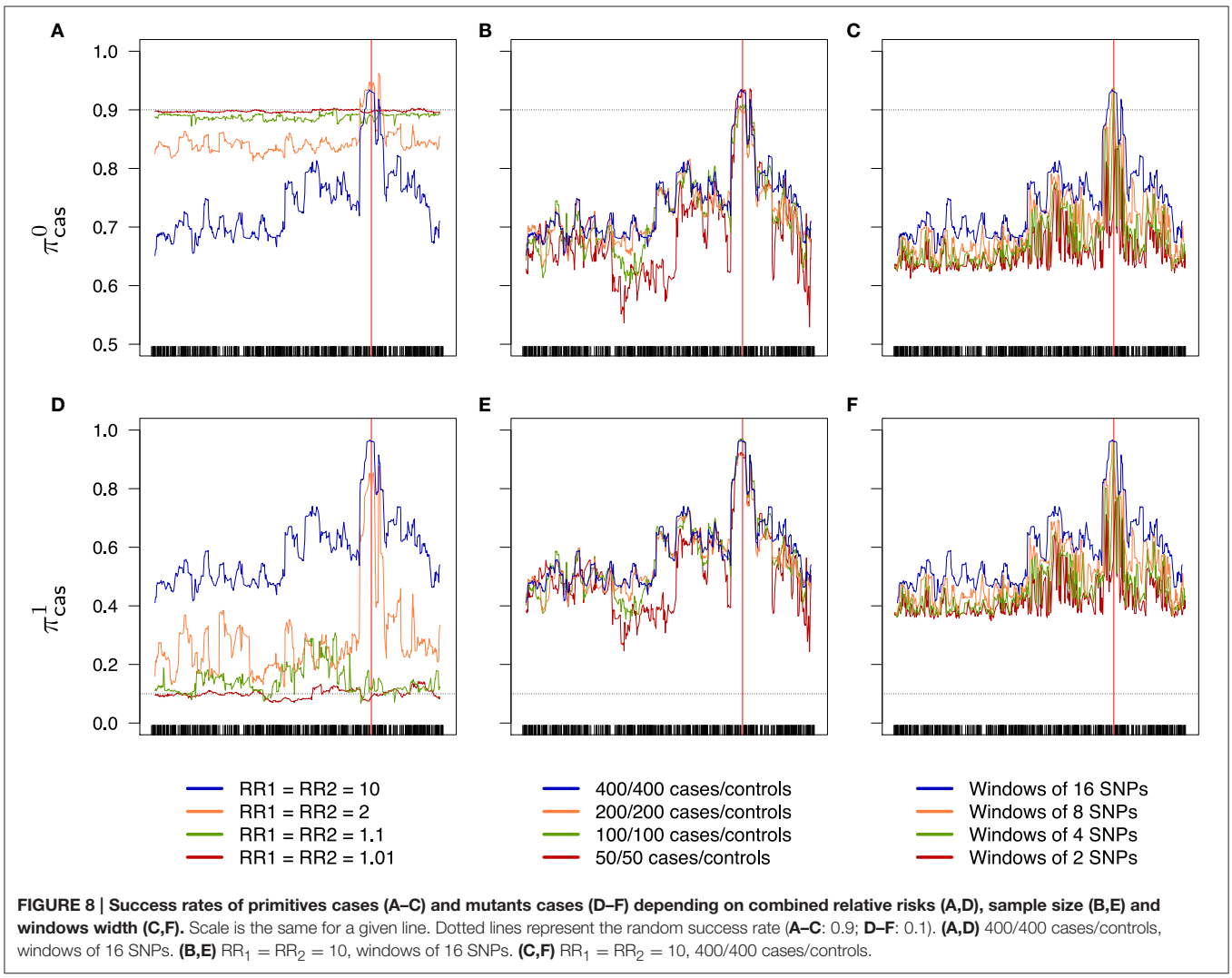

FIGURE 8 | Success rates of primitives cases (A–C) and mutants cases (D–F) depending on combined relative risks (A,D), sample size (B,E) and windows width (C,F). Scale is the same for a given line. Dotted lines represent the random success rate (**A–C**: 0.9; **D–F**: 0.1). **(A,D)** 400/400 cases/controls, windows of 16 SNPs. **(B,E)** $RR_1 = RR_2 = 10$, windows of 16 SNPs. **(C,F)** $RR_1 = RR_2 = 10$, 400/400 cases/controls.

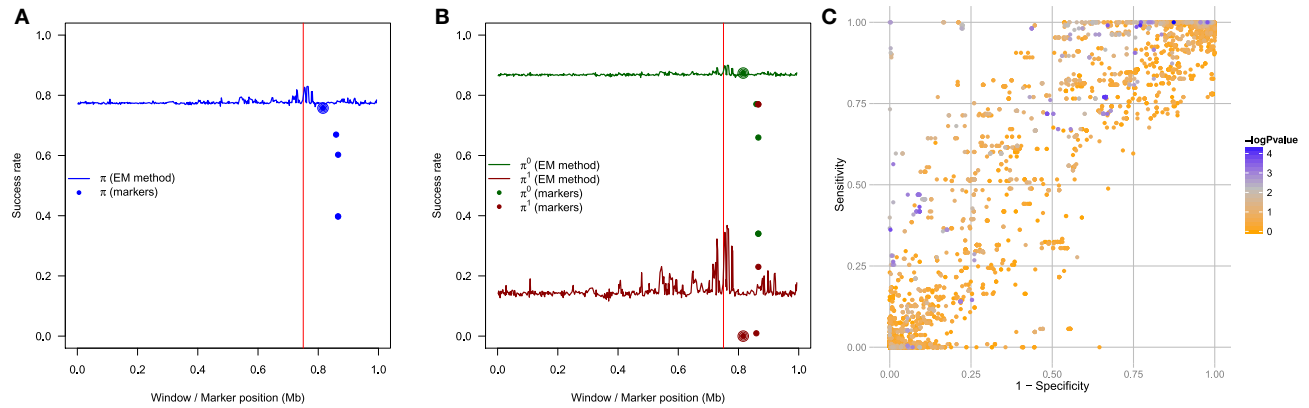

FIGURE 9 | Using the genetic model $RR_1 = RR_2 = 2$. (A) The blue line indicates the global success rate when the TIM is estimated by our proposed methodology, and the four dots around the position of the TIM indicate the success rate when the TIM is estimated by the four most significant markers (the naive method); the marker showing the strongest association is the largest of the four points. **(B)** The green line indicates the success rate π^0 (specificity) and the the red line indicates the success rate π^1 (sensitivity) both obtained with TIM estimated by our proposed methodology; the four green dots are the specificity of the four most associated markers (naive method), and the four red dots are the sensitivity of the same four markers (naive method). **(C)** Relation between sensitivity and 1-specificity, showing $-\log_{10}$ of the p-value for each marker.

position of these markers. One can see that the success rate for the naive method is lower than the success rate of the EM method around the true position of the TIM, which is expected since the EM method benefits from the haplotype and the phenotype information, and from the penetrance model. Another benefit is that the EM method does not need many markers, we used only 500 of them. Moreover, it is very interesting to evaluate the sensitivity π^1 and the specificity π^0 for the naive method, and to compare them with our previous estimates obtained using our method. This comparison is illustrated in **Figure 9B**, which shows π^1 and π^0 (from **Figure 9B**), and the same rates for the four most associated markers using the naive method. The EM method (plain lines) shows both increased sensitivity and increased specificity around the location of the TIM, as expected, whereas the naive method has a very high specificity and a very low sensitivity. To complement these results, **Figure 9C** shows the relation of the sensitivity to 1-specificity, such that the darkness of each of the 7503 points is proportional to $-\log_{10}$ p-value of the association between the marker and the phenotype; one can see that the most significantly associated markers are not the ones exhibiting the highest sensitivity and specificity. The best marker (regarding sensitivity and specificity) is at the top-left of the figure, and is not near being the most associated one. It is important to note, however, that these results depend of the strength of the genetic model: if $RR_1 = RR_2 = 10$, then it is likely that one of the marker could perform as good or better than the EM method, because the association between the marker and the phenotype in this case would be more direct. This illustrates that our EM method surpasses the naive method in the most interesting cases.

5. Conclusion

We have shown how to build an EM algorithm to jointly estimate haplotypes and unknown alleles at the TIM conditionally on the phenotype. In contrast to other methodologies, we use the phenotypic information available, and estimate the frequencies of haplotypes for non carriers, V_0, and for carriers, V_1; the method also estimates the alleles $T = (\delta_1, \delta_2)$, opening new avenues. This method to estimate the alleles at the TIM can also be used with resolved haplotypes, and with missing values. We have shown that the methodology is robust to the sampling from case/control design, which is commonly used in genetic studies. We benchmarked the method on data simulated under the coalescent. The efficiency of the method to infer the alleles at the TIM depends mostly on the strength of the genetic model: when the relative risks are high, the success rates of correctly estimating the alleles are high. This implies that it would be relatively easy to infer TIM alleles for mendelian traits, however we probably need more data when relative risks are low. We observed that neither the frequency of the disease nor of or the causal alleles in the population had any impact on the efficiency of the method. This was to be expected given the case/control design, which implies an enrichment in cases, and thus in causal alleles, in the samples. We also compared our methodology to a naive method, which consists of estimating the alleles at the TIM by the alleles of the marker the most significantly associated with the phenotype. By studying specificity and sensitivity, we have shown that the proposed method provides both higher specificity and higher sensitivity, especially around the true position of the TIM.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada by a grant to the first FL. GB has received scholarships from FQRNT and NSERC.

Acknowledgments

We thank our colleague S. Froda for helpful comments on earlier versions of the manuscript.

References

Boucher, G. (2009). *Intégration de la Réalité Diploïde et des Modèles de Pénétrance à une Méthode de Cartographie Génétique Fine*. Master's thesis, Université du Québec à Montréal.

Browning, S. R., and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 12, 703–714. doi: 10.1038/nrg3054

Clark, A. G. (1990). Inference of haplotypes from pcr-amplified samples of diploid. *Mol. Biol. Evol.* 7, 111–122.

Dupont, M. (2012). *Cartographie Génétique Fine : Évaluation d'une Méthode D'estimation des Allèles et du Modèle de Pénétrance*. Master's thesis, Université du Québec à Montréal.

Excoffier, L., Dupanloup, I., Huerta-Sãnchez, E., Sousa, V. C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9:e1003905. doi: 10.1371/journal.pgen.1003905

Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12, 921–927.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohem, D., et al. (2001). Genetic analysis of case/control data using estimated haplotype frequencies: application to apoe locus variation and alzheimer's disease. *Genome Res.* 11, 143–151. doi: 10.1101/gr.148401

Foulkes, A. S. (2009). *Applied Statistical Genetics with R: for Population-based Association Studies*. New York, NY: Springer.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338. doi: 10.1093/bioinformatics/18.2.337

Larribe, F., Lessard, S., and Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theor. Popul. Biol.* 62, 215–229. doi: 10.1006/tpbi.2002.1601

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., et al. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 78, 437–450. doi: 10.1086/500808

Minichiello, M. J., and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* 79, 910–922. doi: 10.1086/508901

Qin, Z. S. Q., Niu, T., and Liu, J. (2002). Partition-ligation expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 71, 1242–1247. doi: 10.1086/344207

Stephens, M., and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73, 1162–1169. doi: 10.1086/379378

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978–989. doi: 10.1086/319501

Zöllner, S., and Pritchard, J. K. (2005). Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* 169, 1071–1092. doi: 10.1534/genetics.104.031799

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.