BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 206–213
doi:10.1093/bioinformatics/btt609

Genetics and population analysis

Advance Access publication October 24, 2013

Fast pairwise IBD association testing in genome-wide association studies

Buhm Han[1,2,3,4,†], Eun Yong Kang[5,†], Soumya Raychaudhuri[1,2,3,4,6], Paul I. W. de Bakker[1,4,7,8] and Eleazar Eskin[5,9,*]

[1]Division of Genetics, Brigham and Women's Hospital and [2]Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, [3]Partners Center for Personalized Genetic Medicine, Boston, MA 02115, USA, [4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA, [5]Computer Science Department, University of California, Los Angeles, CA 90095, USA, [6]Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK, [7]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands, [8]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands and [9]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Recently, investigators have proposed state-of-the-art Identity-by-descent (IBD) mapping methods to detect IBD segments between purportedly unrelated individuals. The IBD information can then be used for association testing in genetic association studies. One approach for this IBD association testing strategy is to test for excessive IBD between pairs of cases ('pairwise method'). However, this approach is inefficient because it requires a large number of permutations. Moreover, a limited number of permutations define a lower bound for *P*-values, which makes fine-mapping of associated regions difficult because, in practice, a much larger genomic region is implicated than the region that is actually associated.

Results: In this article, we introduce a new pairwise method 'Fast-Pairwise'. Fast-Pairwise uses importance sampling to improve efficiency and enable approximation of extremely small *P*-values. Fast-Pairwise method takes only days to complete a genome-wide scan. In the application to the WTCCC type 1 diabetes data, Fast-Pairwise successfully fine-maps a known human leukocyte antigen gene that is known to cause the disease.

Availability: Fast-Pairwise is publicly available at: http://genetics.cs.ucla.edu/graphibd.

Contact: eeskin@cs.ucla.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 2, 2013; revised on September 7, 2013; accepted on October 21, 2013

1 INTRODUCTION

Identity by descent (IBD) is a fundamental concept in genetics. Two individuals are IBD at a locus if they have identical alleles inherited from a common ancestor. Investigators have put tremendous efforts to map the IBD segments between purportedly unrelated individuals (Browning and Browning, 2010; Gusev *et al*., 2009; Purcell *et al*., 2007). The current state-of-the-art methods such as GERMLINE (Gusev *et al*., 2009) and Beagle (Browning and Browning, 2010, 2011) can detect even small (several megabases) IBD segments shared between individuals.

One promising application of IBD mapping is to use discovered IBD segments in the association testing (Purcell *et al*., 2007). Investigators usually test single nucleotide polymorphisms (SNPs) for association, but SNPs may not 'tag' low frequency causal variations well (de Bakker *et al*., 2005). Imputation also performs poorly on rare variants (Browning and Browning, 2009; Marchini *et al*., 2007). Association testing based on the IBD information, or 'IBD association testing', can complement standard association testing methods (Browning and Browning, 2011).

There are two categories of IBD association testing method. The first method is the pairwise method (Purcell *et al*., 2007), where one compares the IBD rate of case/case pairs with the background IBD rate to detect excessive IBD between cases. The rationale is that if a rare causal variation has occurred in a relatively recent ancestor, cases will likely share an IBD segment containing the causal variant. The second method is the clustering method (Gusev *et al*., 2011), where one divides individuals into clusters based on the IBD information and then test each cluster for association assuming the cluster 'tags' a rare causal variation. In this article, we focus on the pairwise method.

The pairwise method has two computational challenges. The first challenge is computational efficiency. In the pairwise method, one uses permutation to approximate *P*-values because it is difficult to analytically obtain the asymptotic distribution of the statistic. Because the *P*-value threshold for genome-wide association studies (GWAS) is necessarily low due to multiple testing (Browning and Thompson, 2012), one must perform a large number of permutations, which can be computationally demanding. The second challenge is fine-mapping. After one identifies significant loci, it is important to pinpoint the most significant peak within the loci to further follow up candidate genes. The permutation is limited for this purpose because the smallest *P*-value it can approximate is constrained by the number

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, first two authors should be regarded as joint First Authors.

of permutations, often resulting in many SNPs having the same minimal *P*-values in the region.

In this article, we present a new method, 'Fast-Pairwise', to overcome the computational challenges of the traditional pairwise method. Fast-Pairwise uses 'importance sampling' (Wasserman, 2004) to improve efficiency and to enable approximation of extremely small *P*-value. To devise an importance sampling procedure, we introduce a new statistic that has two properties; it can approximate the pairwise method statistic, and it can be conveniently used for designing a sampling procedure. We show that the new statistic has a close relationship with the pairwise method statistic through the properties of the graph representation of IBD.

Fast-Pairwise is efficient and takes only days to complete a genome-wide scan. To demonstrate the utility in fine-mapping, we apply our method to the type 1 diabetes dataset of the Wellcome Trust Case Control Consortium (2007). In this dataset, the traditional pairwise method can identify a significant region in chromosome 6 (Browning and Thompson, 2012), but it gives the same minimal *P*-value for a wide region (26.7–35.5 Mb), including all eight classical human leukocyte antigen (HLA) genes. Among these, Fast-Pairwise pinpoints *HLA–DQB1*, which is known to cause the disease (Todd *et al.*, 1987).

Fast-Pairwise is publicly available at http://genetics.cs.ucla.edu/graphibd.

2 METHODS

2.1 IBD graph

Given N individuals, the IBD information at a genomic locus can be represented as a graph with N vertices (Fig. 1). An edge exists between a pair of vertices if the individuals are IBD.

2.2 Pairwise methods for IBD association mapping

We refer to a class of IBD association mapping methods as 'pairwise methods' if they examine the relative number of edges in the IBD graph at each locus. There are three different types of edges: edges that connect two case individuals, edges that connect two control individuals and edges that connect a case and a control individual. Pairwise methods can be performed in two different ways. One way is to compare the number of case/case pairs with control/control pairs. A second way is to compare the number of case/case pairs with non-case/case pairs (union of control/control pairs and case/control pairs). We will call the first method CC and the second method CN. In this article, we mainly focus on CN consistent with previous studies (Browning and Thompson, 2012; Purcell *et al.*, 2007). If we simply refer to the 'pairwise method', we are referring to CN.

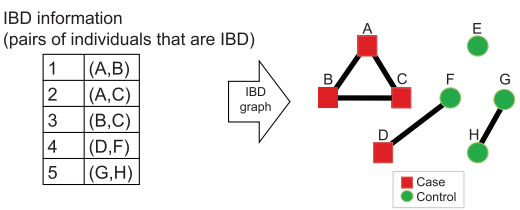

Fig. 1. An example of IBD graph. IBD detection method provides IBD information (Table). Then we build a graph where vertices are individuals and edges are IBD relationships

Suppose that we have N^+ cases and N^- controls ($N^+ + N^- = N$). Let V^+ be the set of case vertices and V^- be the set of control vertices. Let E^{++} be the set of all possible case/case vertex pairs, E^{--} be the set of all possible control/control vertex pairs and E^{+-} be the set of all possible case/control vertex pairs. Let e_{ij} be 1 if there exists an edge between vertices i and j, and 0 otherwise. Then, the CC and CN statistics are defined as

$$S_{CC} = \widehat{IBD}_{\text{case/case}} - \widehat{IBD}_{\text{control/control}} = \sum_{(i,j)\in E^{++}} \frac{e_{ij}}{\binom{N^+}{2}} - \sum_{(i,j)\in E^{--}} \frac{e_{ij}}{\binom{N^-}{2}}$$

$$S_{CN} = \widehat{IBD}_{\text{case/case}} - \widehat{IBD}_{\text{non-case/case}} = \sum_{(i,j)\in E^{++}} \frac{e_{ij}}{\binom{N^+}{2}} - \sum_{(i,j)\in E^{--}\cup E^{+-}} \frac{e_{ij}}{\binom{N^-}{2} + N^+N^-}$$

The asymptotic distributions of these statistics are difficult to obtain analytically. This is because the statistics are based on the edges that depend on each other. For this reason, statistical significance is assessed by permutation. We assume a one-sided test, where IBD segments carry variants that are involved in disease (Browning and Thompson, 2012).

The relationship between CC and CN is worth noting. Under the condition that the background IBD rates of control/control pairs and the non-case/case pairs are equivalent ($IBD_{\text{control/control}} = IBD_{\text{non-case/case}}$), CN will be more powerful than CC owing to the additional N^+N^- pairs it considers. We expect that, however, the relative ordering of the two statistics is similar to each other (Fig. 2), as most of the pairs for both CC and CN are the same. As we will show later, we will use this similarity as the basis of increasing the computational efficiency of computing the significance of S_{CN}.

2.3 Permutation test

Permutation is the standard approach for assessing the significance of the pairwise method. A single permutation can be thought of as a randomly sampled a vector of case/control disease statuses. Let

$$\mathbf{v} = (v_1, \ldots, v_N), \quad \forall v_i \in \{0, 1\}$$

be the vector of disease status of N individuals, where 0 denotes control and 1 denotes case. The test statistic of pairwise method, S_{CN}, is a function of $\mathbf{v}$. Let $\hat{\mathbf{v}}$ be the case/control status that was originally observed in the data. The standard permutation test is equivalent to sampling new $\mathbf{v}$ from all possible permutations of $\hat{\mathbf{v}}$ assuming a uniform distribution. Let $\mathbf{B}$ be the set of sampled $\mathbf{v}$. The estimated *P*-value is

$$\hat{p} = \frac{1}{|\mathbf{B}|} \sum_{\mathbf{v}\in\mathbf{B}} \delta(S_{CN}(\mathbf{v}) \geq S_{CN}(\hat{\mathbf{v}})) \quad (1)$$

where δ is the indicator function.

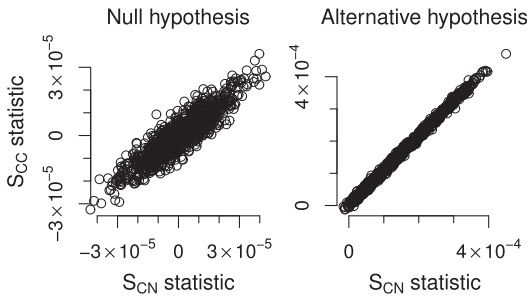

Fig. 2. High correlation between CC and CN statistics. We simulated 1000 studies under the alternative hypothesis (see Section 3.2). We then permuted phenotypes to simulate the null hypothesis. Spearman ρ is 0.89 and 0.99 under the null and the alternative, respectively

The drawback of the permutation test is that it is computationally inefficient. If the true *P*-value is small, which is required in genome-wide studies, we will need a large number of permutations. For example, to assess a *P*-value p with standard error $p/10$, one needs approximately $100/p$ samples. For the genome-wide threshold of IBD association testing [6×10^{-6}, Browning and Thompson (2012)], >10 million permutations are required.

2.4 Fast-Pairwise

We develop a new method 'Fast-Pairwise' that uses importance sampling technique to speed up CN method (Wasserman, 2004). Unlike the standard permutation test that samples case/control status $\mathbf{v}$ from the uniform distribution, we sample $\mathbf{v}$ non-uniformly. Specifically, we aim to sample $\mathbf{v}$ from all permutations of $\hat{\mathbf{v}}$ such that, on average, $S_{CN}(\mathbf{v})$ will be similar to $S_{CN}(\hat{\mathbf{v}})$. The intuition is that by intentionally sampling $\mathbf{v}$ that gives large value of S_{CN}, we can reduce the variance of the *P*-value estimate. Thus, our goal is to design a sampling procedure that satisfies

$$\mathbb{E}_f(S_{CN}(\mathbf{v})) = S_{CN}(\hat{\mathbf{v}}) \tag{2}$$

where the expectation is with respect to f, our sampling distribution for $\mathbf{v}$. However, designing such a sampling procedure is not straightforward. To this end, we leverage the fact that we can construct a simpler statistic that approximates S_{CN}, which we use for the sampling.

2.5 IBD-degreetype

To apply importance sampling, we must identify a statistic that roughly approximates S_{CN} but can be conveniently used for designing a sampling procedure. Because we empirically have observed that S_{CC} approximates S_{CN} (Fig. 2), we want to find a statistic that approximates S_{CC}. Our proposed statistic S_{SUM} is related to S_{CC} through a concept that we introduce, called the IBD-degreetype, which is simply the degree of each individual in the IBD graph. Obtaining the degrees of vertices is equivalent to splitting all edges and counting how many split edges are adjacent to each vertex (Fig. 3). Then we assign these numbers to the vertices. Given this, we define the IBD-degreetype as conceptually similar to a genotype where the allele of each individual equals to the degree of the corresponding vertex in the IBD graph.

The IBD-degreetypes can be used for statistical testing for IBD association testing. The intuition is that if case/case pairs have an excessive number of IBDs, then case vertices will have higher degrees than control vertices. The test based on IBD-degreetype will be comparing the average degrees between cases and controls,

$$S_{ID} = \sum_{i \in V^+} \frac{w_i}{N^+} - \sum_{i \in V^-} \frac{w_i}{N^-} \tag{3}$$

where w_i is the IBD-degreetype of individual i, or equivalently the degree of vertex i in the graph. We note that this statistic is conceptually similar to the traditional association statistic, which compares the frequency of the genotypes between the cases and controls. Here we instead compute the difference between the case and controls of the IBD-degreetypes (hence the name).

We are interested in the monotonic relationship between statistics within the permutation procedure. Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be the permuted case/control status. We define 'monotonic increasing relationship' (MIR) as follows.

DEFINITION 1. Two statistics S and T are in an MIR if, $\forall \mathbf{v}_1 \neq \mathbf{v}_2$, $S(\mathbf{v}_1) \geq S(\mathbf{v}_2)$ *iff* $T(\mathbf{v}_1) \geq T(\mathbf{v}_2)$.

It is clear that if two statistics are in MIR, they will give the same *P*-value under permutation. It also follows that MIR is transitive (if S and T are in MIR and T and U are in MIR, then S and U are in MIR).

The IBD-degreetype test has the following relationship to the pairwise CC method. Figure 3 illustrates this relationship with a toy example.

LEMMA 1. In a balanced study design ($N^+ = N^-$), CC and the IBD-degreetype test are in MIR.

PROOF.

$$\begin{aligned}
S_{ID} &= \sum_{i \in V^+} \frac{w_i}{N^+} - \sum_{i \in V^-} \frac{w_i}{N^-} = \frac{1}{N^+}\left\{\sum_{i \in V^+} w_i - \sum_{i \in V^-} w_i\right\} \\
&= \frac{1}{N^+}\left\{\left(2\sum_{(i,j)\in E^{++}} e_{ij} + \sum_{(i,j)\in E^{+-}} e_{ij}\right) - \left(2\sum_{(i,j)\in E^{--}} e_{ij} + \sum_{(i,j)\in E^{+-}} e_{ij}\right)\right\} \\
&= \frac{2}{N^+}\left\{\sum_{(i,j)\in E^{++}} e_{ij} - \sum_{(i,j)\in E^{--}} e_{ij}\right\} \\
&= \frac{2}{N^+}\binom{N^+}{2}\left\{\sum_{(i,j)\in E^{++}} \frac{e_{ij}}{\binom{N^+}{2}} - \sum_{(i,j)\in E^{--}} \frac{e_{ij}}{\binom{N^-}{2}}\right\} \\
&= \frac{2}{N^+}\binom{N^+}{2} S_{CC} = (N^+ - 1)\, S_{CC}
\end{aligned}$$

Because S_{ID} and S_{CC} differ by a constant multiplication factor $(N^+ - 1)$, they are in MIR.

We introduce another simple form of test statistic based on IBD-degreetype. This *sum statistic* is the sum of IBD-degreetype alleles or the degrees of the vertices in cases,

$$S_{SUM} = \sum_{i \in V^+} w_i$$

LEMMA 2. S_{ID} and S_{SUM} are in MIR.

PROOF. Note that

$$\sum_{(i,j)\in E^{++}} e_{ij} + \sum_{(i,j)\in E^{+-}} e_{ij} + \sum_{(i,j)\in E^{--}} e_{ij} = |e|$$

where $|e|$ denotes the total count of edges. In addition, the sum of the IBD-degreetypes over all vertices is equal to $2|e|$ because each edge is counted twice.

$$\sum_{i \in V^+} w_i + \sum_{i \in V^-} w_i = 2|e|$$

Therefore,

$$\begin{aligned}
S_{ID} &= \sum_{i \in V^+} \frac{w_i}{N^+} - \sum_{i \in V^-} \frac{w_i}{N^-} = \left(\frac{1}{N^+} + \frac{1}{N^-}\right)\sum_{i \in V^+} w_i - \frac{2|e|}{N^-} \\
&= \left(\frac{1}{N^+} + \frac{1}{N^-}\right) S_{SUM} - \frac{2|e|}{N^-}
\end{aligned}$$

Because $\left(\frac{1}{N^+} + \frac{1}{N^-}\right) > 0$ and $\frac{2|e|}{N^-}$ is a constant, S_{SUM} is a monotonic increasing linear transformation of S_{ID}. Thus, they are in MIR.

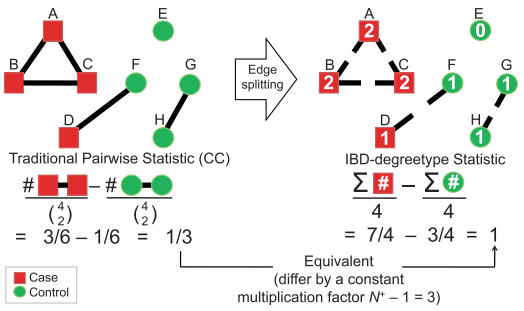

Fig. 3. Equivalence of pairwise (CC) method and IBD-degreetype test in a balanced study. Note that in pairwise CC method, the edge between (D,F) pair is ignored, which would not be ignored in CN method

2.6 Substitution strategy

Here we propose to use S_{SUM} in sampling as a substitution to the pairwise method statistic, S_{CN}. The logical ground for this strategy comes from the relationship between S_{CN} and S_{SUM}. We have empirically shown that S_{CN} and S_{CC} are highly correlated (Fig. 2). Because S_{CC} and S_{ID} are in MIR in a balanced study (Lemma 1), we expect that they will be correlated in general even in an unbalanced study, and we show this property through a simulation experiment (Supplementary Fig. S1). S_{ID} is in MIR with S_{SUM} (Lemma 2). Thus, S_{SUM} can be an approximation to S_{CN} (Fig. 4).

Given this relationship between S_{CN} and S_{SUM}, our strategy is to sample $\mathbf{v}$ such that $S_{SUM}(\mathbf{v})$ will be similar to $S_{SUM}(\hat{\mathbf{v}})$ on average. Our new goal can be described as

$$\mathbb{E}_f(S_{SUM}(\mathbf{v})) = S_{SUM}(\hat{\mathbf{v}}) \tag{4}$$

It turns out that this new goal is much easier to achieve. Note that S_{SUM} is used only for sampling. After the sampling is done, the sampled $\mathbf{v}$ is used to approximate the P-value of CN method.

This substitution approach works because in importance sampling, the sampling distribution need not guarantee optimality (the smallest variance of P-value estimate). Instead, a reasonably similar distribution to the optimal distribution suffices. It is clear that this strategy will perform the best if the balanced study condition is met. However, even if the condition is not met, only the variance of P-value estimate is affected and not the mean. The P-value estimate will still be unbiased, and it only means that we will need a larger number of samples to obtain the same accuracy.

2.7 Sampling with replacement

In this section, we devise a sampling procedure satisfying Equation (4). Such a sampling procedure will be the core part in our importance sampling framework for speeding up CN method.

Sampling a random $\mathbf{v}$ from all permutations of $\hat{\mathbf{v}}$ can be thought of as sampling N^+ of N individuals that will be assigned case status, or equivalently, sampling N^+ case indices among $1, \ldots, N$. Let $a_1, \ldots, a_{N^+}$ be the sampled case indices. These are the indices in $\mathbf{v}$ that will be assigned '1'. Sampling case indices is without-replacement sampling procedure; we cannot sample the same index twice, so that exactly N^+ distinct indices will be sampled at the end ($\forall i \neq j, a_i \neq a_j$). This way, we can restrict sample space of $\mathbf{v}$ to the set of all permutations of $\hat{\mathbf{v}}$.

However, the design of sampling procedure satisfying Equation (4) is considerably easier if we assume sampling case indices with replacement. That is, we allow the same index can be sampled multiple times. Although this assumption is not valid for our purpose, our strategy is to devise a sampling approach satisfying Equation (4), assuming sampling with replacement first, and then extend the approach to the sampling without replacement.

Suppose that we pick a_1 among $1, \ldots, N$ with probability $P(a_1 = k) = g(k)$, $\sum_{k=1}^{N} g(k) = 1$. Because we assume sampling with replacement, sampling a_2 is no different from sampling a_1; in fact, for any $1 \leq i \leq N^+$, a_i is independent and identically distributed (IID) with distribution g. Now consider w_{a_1}, the IBD-degreetype allele of a_1. Let

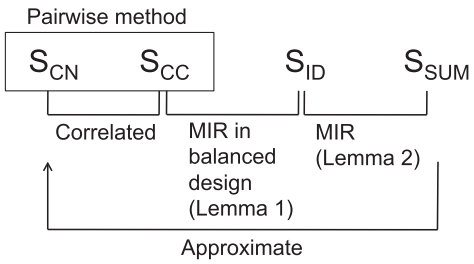

Fig. 4. Relationship between different statistics

$\mathbb{E}_g(w_{a_1})$ be the expected value of w_{a_1} with respect to g. Again, because we assume sampling with replacement, $\mathbb{E}_g(w_{a_1}) = \ldots = \mathbb{E}_g(w_{a_{N^+}})$.

Then, by the definition of S_{SUM}, we can easily see that

$$\mathbb{E}_g(S_{SUM}) = N^+ \mathbb{E}_g(w_{a_1})$$

Thus, equation (4) can be described as

$$N^+ \mathbb{E}_g(w_{a_1}) = S_{SUM}(\hat{\mathbf{v}})$$

or

$$\mathbb{E}_g(w_{a_1}) = \frac{S_{SUM}(\hat{\mathbf{v}})}{N^+} \tag{5}$$

where the left side is the expected case IBD-degreetype allele in our distribution g and the right side is the average case IBD-degreetype allele in the observation $\hat{\mathbf{v}}$. This shows that, if the P-value is highly significant (e.g. the right side is large), we should pick a_1 (and all a_i) such that the expected value of IBD-degreetype allele can be large.

Here we propose a new sampling procedure that satisfies condition (5). We define distribution g as follows

$$P(a_1 = k) = g(k) \propto t_k \quad \text{where} \quad t_k = 1 + \rho w_k$$

and

$$\rho = \frac{N \frac{S_{SUM}(\hat{\mathbf{v}})}{N^+} - \sum_{k=1}^{N} w_k}{\sum_{k=1}^{N} w_k^2 - \frac{S_{SUM}(\hat{\mathbf{v}})}{N^+} \sum_{k=1}^{N} w_k} \tag{6}$$

It is easy to show that this sampling procedure meets condition (5), as condition (5) can be described as

$$\mathbb{E}_g(w_{a_1}) = \frac{\sum_{k=1}^{N} t_k w_k}{\sum_{k=1}^{N} t_k} = \frac{\sum_{k=1}^{N} (1 + \rho w_k) w_k}{\sum_{k=1}^{N} (1 + \rho w_k)} = \frac{S_{SUM}(\hat{\mathbf{v}})}{N^+}$$

and by solving this for ρ, we exactly have equation (6). If $\rho < 0$, then we set ρ to 0 to prevent negative t_k. Such a case is not of our interest at any rate, as we focus on the one-sided test for detecting excess of case/case IBD. We choose the most simple linear function for t_k, which enables us to calculate ρ easily. As a result, for any $\hat{\mathbf{v}}$, we can calculate ρ and completely define the distribution g. So far, we have successfully defined a sampling procedure satisfying Equation (4), assuming sampling with replacement.

2.8 Sampling without replacement

In this section, we extend the sampling procedure from the previous section to the 'sampling without replacement'. Here we propose to heuristically apply the same sampling scheme based on $t_1, \ldots, t_N$ to the without-replacement context. When we pick a_i (ith case index), we pick index k among $\{1, \ldots, N\} \backslash \{a_1, \ldots, a_{i-1}\}$ with probability $t_k / \left(\sum_{j=1}^{N} t_j - \sum_{l=1}^{i-1} t_{a_l} \right)$. That is, we assume the same sampling probability proportional to t_k, but we exclude indices previously picked as cases from our consideration.

However, this sampling procedure does not exactly satisfy Equation (4) in the without-replacement sampling context. The indices with larger IBD-degreetype alleles are likely to be picked as cases earlier in the procedure and removed. Thus, if we use ρ calculated assuming sampling with replacement, the expected case IBD-degreetype allele [the left side of Equation (5)] will be smaller than what we would obtain in the with-replacement context. To compensate for this difference, we use the following heuristic. In the middle of sampling, we empirically assess the left side of Equation (4) by examining the currently obtained samples. Then,

we dynamically increment ρ until the left side of Equation (4) is close to the right side. This simple heuristic is sufficient because, again, in importance sampling, we only need to approximately satisfy Equation (4).

2.9 *P*-value calculation

By using the sampling procedure developed in the previous section, we can obtain many sample $\mathbf{v}$ that approximately satisfy Equation (2). The final step is to use these samples to assess the *P*-value of pairwise (CN) method. In importance sampling, we must account for the fact that we used a sampling distribution that is different from the original distribution. Our original distribution is the uniform distribution defined by the standard permutation approach. The sampling distribution is defined by the sampling procedure that we developed in the previous section.

For a given $\mathbf{v}$, the probability of sampling $\mathbf{v}$ differs in the two distributions as follows. To sample a $\mathbf{v}$, we sample the case indices $a_1, \ldots, a_{N^+}$. The probability of sampling case indices in the standard uniform distribution is

$$P_{\text{Uniform}}(\mathbf{v}) = \frac{1}{N \cdots (N - N^+ + 1)}$$

On the other hand, the probability of sampling case indices in our sampling procedure described in the previous section is

$$P_{\text{New}}(\mathbf{v}) = \prod_{k=1}^{N^+} \left\{ t_{a_k} \Big/ \left(\sum_{j=1}^{N} t_j - \sum_{l=1}^{k-1} t_{a_l} \right) \right\}$$

This is because at each step we pick ith case index a_i, we sample the index with probability proportional to t_{a_i}, where the previously picked indices $a_1, \ldots, a_{N^+}$ are excluded from consideration. Using the standard formula of importance sampling, we approximate the *P*-value

$$\hat{p} = \frac{1}{|\mathbf{B}|} \sum_{\mathbf{v} \in \mathbf{B}} \frac{P_{\text{Uniform}}(\mathbf{v})}{P_{\text{New}}(\mathbf{v})} \cdot \delta(S_{CN}(\mathbf{v}) \geq S_{CN}(\hat{\mathbf{v}}))$$

Note that we use the pairwise CN statistic in this formula. We use S_{SUM} only to facilitate the sampling of $\mathbf{v}$, and then use the obtained samples for CN method at the final step.

We can approximate the variance of $\hat{p}$ with the following formula

$$\left\{ \frac{1}{|\mathbf{B}|} \sum_{\mathbf{v} \in \mathbf{B}} \left(\frac{P_{\text{Uniform}}(\mathbf{v})}{P_{\text{New}}(\mathbf{v})} \right)^2 \cdot \delta(S_{CN}(\mathbf{v}) \geq S_{CN}(\hat{\mathbf{v}})) - \hat{p}^2 \right\} \Big/ |\mathbf{B}|$$

2.10 Adjusting for population structure

A simple correction for population structure has been previously proposed (Browning and Thompson, 2012; Purcell *et al.*, 2007) for the pairwise method. In this simple approach, the genomic average is subtracted from each of the two contrasting terms of the statistic. For example, in CN method, the genomic average of case/case IBD rate is subtracted from the observed case/case IBD rate, and the genomic average of non-case/case IBD rate is subtracted from the observed non-case/case IBD rate before calculating the statistic. The same approach can be applied to our Fast-Pairwise method.

3 RESULTS

3.1 Efficiency

To assess the efficiency gain of our Fast-Pairwise method, we use the Wellcome Trust Case Control Consortium (WTCCC) data (Wellcome Trust Case Control Consortium, 2007). We first run Beagle FastIBD to detect IBD between individuals (Browning and Browning, 2011). Then we test individual IBD segments for associations. We perform IBD association testing using both the traditional pairwise method based on permutation and our Fast-Pairwise method. We perform 10 million permutations for the traditional pairwise method. For our Fast-Pairwise method, we perform importance sampling with 1000 samples and 10 000 samples. We implemented both methods in the Java programming language.

Table 1 shows the estimated running time of both methods for analyzing the whole genome data (500 000 SNPs) of a single disease. The time is extrapolated from the estimated time for chromosome 22. Our Fast-Pairwise method is several orders of magnitude faster than the traditional pairwise method. It takes only 4 days for the whole genome, whereas the traditional method can take 13 000 days or 35 years of CPU time.

We can reduce the computation time for the traditional pairwise method by using an adaptive permutation approach. We can terminate the permutation earlier if the *P*-value approximates to a non-significant value. Given a *P*-value p, we need $100/p$ permutations to obtain the standard error of $\sim p/10$. Suppose that we sample $100/p$ permutations for each *P*-value with upper limit of 10 millions. When we apply this adaptive approach to the WTCCC type 1 diabetes data, the estimated computation time is 474 days. Thus, Fast-Pairwise is still an order of magnitude faster than the traditional pairwise method with an adaptive permutation approach.

3.2 Accuracy

To assess the accuracy of our importance sampling, we use the simulation framework similar to Browning and Thompson (2012). Using the HapMap ENCODE regions (International HapMap Consortium, 2005), we run HapGen2 to simulate 10 000 individuals (Su *et al.*, 2011). These individuals define our founder population. Then we simulate the first generation by sampling 100 000 individuals from the founders. Next we simulate the second generation by sampling 100 000 individuals from the first generation. We repeat until we obtain the 25th generation. Finally, we use the 25th generation to simulate a case/control study. Within the ENCODE region, we randomly select five causal variants among all rare variants (minor allele frequency <1%). If a haplotype contains one or more causal variants, it confers risk with relative risk selected from uniform (3,10). We assume the disease prevalence of 0.1. Given this disease model, using the standard formula of the case and control minor allele frequencies (Han *et al.*, 2009), we sample 1500 cases and 1500 controls from the 25th generation. The IBD information between a pair of individuals is determined by tracking whether they are descendants of the same founder. We repeat

Table 1. Running time for pairwise IBD association testing for the WTCCC whole genome data

Traditional pairwise method		Fast-Pairwise	
10^7 permutations	Adaptive permutations	10^4 samples	10^3 samples
35 years	474 days	40 days	4 days

this simulation 100 times per each of the 10 ENCODE regions to generate 1000 sets of case/control studies.

Given these case/control study sets, we assess the *P*-values of pairwise (CN) method using both the standard permutation and the importance sampling of Fast-Pairwise. We use 10 000 samples for importance sampling and compare with 10^4, 10^5 and 10^6 permutations. Figure 5 shows that the *P*-values of two methods track well within the *P*-value range that permutation can approximate (up to *P*-values of 10^{-4}, 10^{-5} and 10^{-6}, respectively). Within this range, the Pearson correlation r^2 of two log *P*-values are 0.98, 0.94 and 0.99, respectively. This shows that our importance sampling procedure obtains accurate *P*-values.

Moreover, Figure 5 emphasizes a fundamental difference between the two methods. In permutation, the range of *P*-values one can obtain is limited by the number of permutations. Given $|B|$ permutations, if none of the permutations exceeds the observed statistic, a conservative approximation of *P*-value is $1/|B|$. In contrast, in Fast-Pairwise, the *P*-value range is not bounded by the number of samples. With a relatively small number of samples (10 000), Fast-Pairwise can obtain accurate *P*-values comparable with the permutation test for a wide range of *P*-values. We also performed extra permutations ($>10^6$) to estimate *P*-values between 10^{-6} and 10^{-8}. The *P*-values of the two methods are still consistent within this *P*-value range (triangles in Fig. 5C).

3.3 Application to WTCCC type 1 diabetes data

Browning and Thompson (2012) applied the pairwise (CN) method to the WTCCC type 1 diabetes (T1D) data based on the Beagle FastIBD IBD mapping results (Browning and Browning, 2011). Using 5 million permutations, they found that the major histocompatibility complex (MHC) region in the chromosome 6 is statistically significant given the genome-wide threshold 6×10^{-6}. Because the MHC association to T1D has been historically known (Todd *et al*., 1987), this result was a validation that IBD association testing can detect the true association signal.

The limitation of Browning and Thompson's (2012) permutation approach is that although it is possible to determine whether each test is significant ($p < 6 \times 10^{-6}$) using 5 million permutations, it is not possible to approximate much smaller *P*-values. Given 5 million permutations, the smallest *P*-value one can approximate is bounded to 2×10^{-7}. In the MHC region, because of the strong signal and the long linkage disequilibrium, the location of the top peak of *P*-value is important for interpreting and fine-mapping the results. Figure 6A shows that the top peak of *P*-value is stretched over a wide region (>8 Mb) including all eight classical HLA genes. Therefore, it is difficult to interpret which HLA gene is likely to be involved in the association.

We applied our Fast-Pairwise to the same dataset. Because Fast-Pairwise is the same pairwise method with increased efficiency, we expected to see the similar results as Browning and Thompson (2012). We discovered significant associations within the MHC region. However, the difference is that because our method can approximate small *P*-values well beyond the genome-wide threshold, it is possible to localize the statistical signal to a single marker (Fig. 6B). The top hit is at SNP rs241432 ($p = 7 \times 10^{-45}$) at the intron of *TAP2* gene. Among

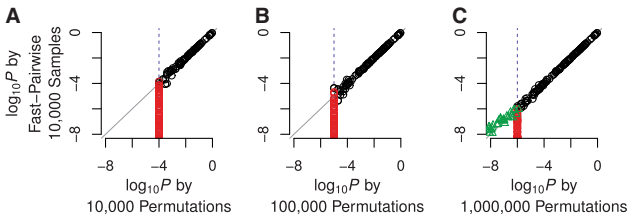

Fig. 5. Accuracy of importance sampling. In simulations using the HapMap ENCODE region, we assess the *P*-values of pairwise (CN) method using both the standard permutation and Fast-Pairwise (importance sampling). We compare 10 000 samples of Fast-Pairwise to (**A**) 10 000, (**B**) 100 000, and (**C**) 1 000 000 permutations. The vertical dashed line denotes the lower bound of *P*-value that permutation can approximate given the number of permutations. The dots along the vertical line denote the simulations where none of the permutations exceeds the observed statistic, and therefore the lower bound of *P*-value is reported by the permutation test. The triangles in (**C**) are the *P*-values that we performed extra permutations ($>10^6$)

all major class I and II HLA genes, the closest gene to this peak is *HLA–DQB1* (150 kb upstream from the peak). It is historically known that the main contributing gene for the MHC association to T1D is *HLA–DQB1* (Todd *et al*., 1987). Thus, this result demonstrates that our Fast-Pairwise method can pinpoint the causal gene among many HLA genes within the MHC region, while the traditional pairwise method cannot.

One interesting observation is that the peak association of our IBD association test is on the *TAP2* gene that encodes antigen peptide transporter 2 and has been shown to confer independent risk to the T1D when conditioned on the DQ haplotypes (Qu *et al*., 2007). Thus, the peak revealed by our Fast-Pairwise method may imply the added effect of *TAP2* in addition to the primary effect of *HLA–DQB1*, which is in linkage disequilibrium.

4 DISCUSSION

We have developed a new efficient method for pairwise IBD association testing called 'Fast-Pairwise'. Fast-Pairwise uses importance sampling and can perform the pairwise method more efficiently than the traditional method based on permutation. Moreover, unlike permutation, Fast-Pairwise can approximate extremely small *P*-values beyond the genome-wide threshold. Using the WTCCC type 1 diabetes data, we show that Fast-Pairwise can successfully pinpoint a gene known to be associated to the disease within the MHC region.

The true utility of the IBD association testing is on finding novel loci where there are potentially multiple rare variants that cannot be found using single SNP tests (Browning and Thompson, 2012). An important advantage of IBD association testing is its wide applicability. The analysis can be performed using the same genotype data collected for single SNP tests without incurring additional cost. For this reason, we feel that many investigators will apply our method to search for these additional loci bearing rare causal variants. What is preventing researchers from applying this approach is an efficient method for IBD association testing, which we provide in this article. We expect that

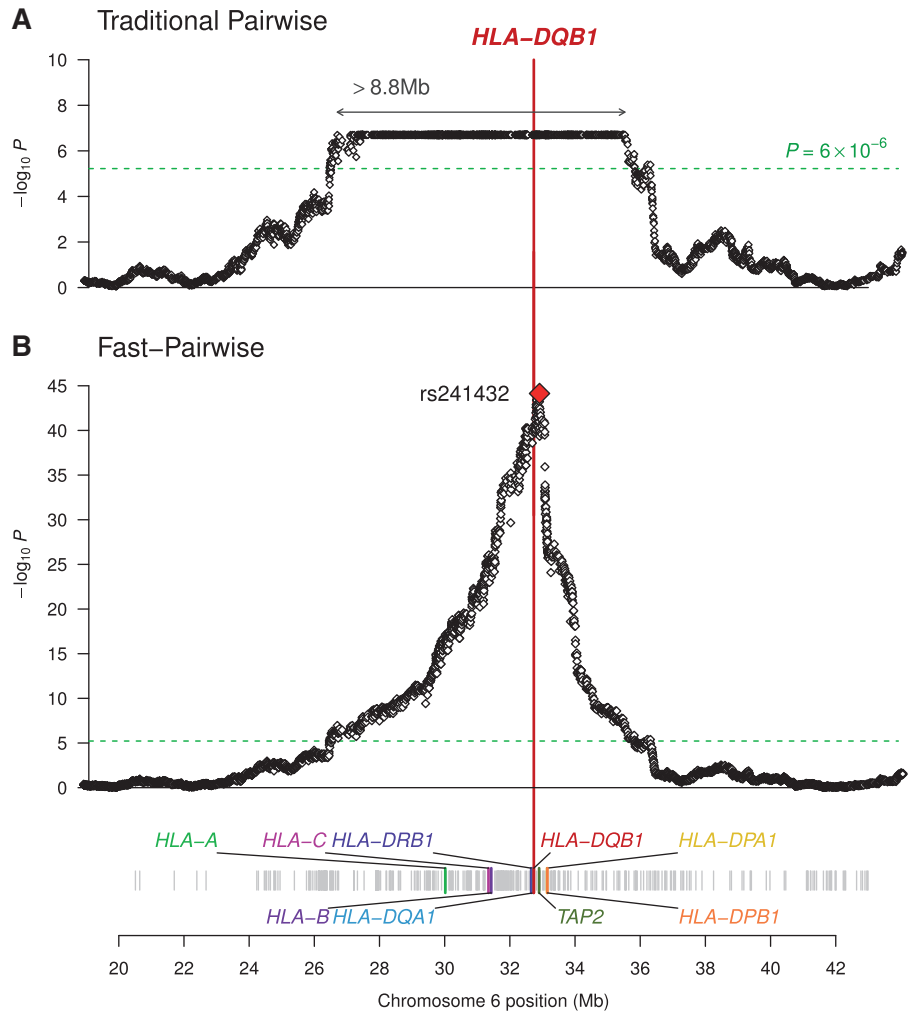

Fig. 6. IBD association testing results of the WTCCC type 1 diabetes data. (**A**) The results that we would obtain if we use the traditional pairwise method. Given 5 million permutations, the smallest *P*-value is bounded to 2×10^{-7}. The top peak is stretched over >8 Mb region complicating the fine-mapping. (**B**) Fast-Pairwise results. The top peak is at *TAP2*, whose closest HLA gene is *HLA-DQB1*. In both plots, the dashed horizontal line denotes genome-wide threshold 6×10^{-6}

our new method will promote the wide use of IBD association testing and facilitate further research on the power and utilities of IBD association testing.

ACKNOWLEDGEMENT

The authors thank the Wellcome Trust Case Control Consortium for data access.

Funding: B.H., E.Y.K. and E.E. are supported by National Science Foundation grants [0513612, 0731455, 0729049, 0916676 and 1065276] and National Institutes of Health grants [K25-HL080079, U01-DA024417, P01-HL30568 and PO1-HL28481]. B.H. is supported by the Samsung Scholarship. B.H., S.R. and P.I.W.d.B are supported by National Institutes of Health grant [NIH-NIAMS 1R01AR062886-01]. P.I.W.d.B. is the recipient of a VIDI award from The Netherlands Organisation for Scientific Research (NWO).

Conflict of Interest: none declared.

REFERENCES

Browning,B.L. and Browning,S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.

Browning,B.L. and Browning,S.R. (2011) A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.*, **88**, 173–182.

Browning,S.R. and Browning,B.L. (2010) High-resolution detection of identity by descent in unrelated individuals. *Am. J. Hum. Genet.*, **86**, 526–539.

Browning,S.R. and Thompson,E.A. (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, **190**, 1521–1531.

de Bakker,P.I.W. *et al.* (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.

Gusev,A. *et al.* (2011) Dash: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am. J. Hum. Genet.*, **88**, 706–717.

Gusev,A. *et al.* (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, **19**, 318–326.

Han,B. *et al.* (2009) Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet.*, **5**, e1000456.
International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
Marchini,J. *et al.* (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
Qu,H.Q.Q. *et al.* (2007) Genetic control of alternative splicing in the tap2 gene: possible implication in the genetics of type 1 diabetes. *Diabetes*, **56**, 270–275.
Su,Z. *et al.* (2011) Hapgen2: simulation of multiple disease SNPs. *Bioinformatics*, **27**, 2304–2305.
Todd,J.A. *et al.* (1987) Hla-dq beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature*, **329**, 599–604.
Wasserman,L. (2004) *All of Statistics*. Springer, New York.
Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.