BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 22 2010, pages 2856–2862
doi:10.1093/bioinformatics/btq529

Genetics and population analysis

Advance Access publication September 24, 2010

RAPID detection of gene–gene interactions in genome-wide association studies

Dumitru Brinza[1], Matthew Schultz[2], Glenn Tesler[3] and Vineet Bafna[4,*]

[1]Life Technologies, Foster City, CA, [2]Graduate Bioinformatics Program, [3]Department of Mathematics and [4]Department of Computer Science and Engineering, Institute for Genomic Medicine, University of California, San Diego, CA, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: In complex disorders, independently evolving locus pairs might interact to confer disease susceptibility, with only a modest effect at each locus. With genome-wide association studies on large cohorts, testing all pairs for interaction confers a heavy computational burden, and a loss of power due to large Bonferroni-like corrections. Correspondingly, limiting the tests to pairs that show marginal effect at either locus, also has reduced power. Here, we describe an algorithm that discovers interacting locus pairs without explicitly testing all pairs, or requiring a marginal effect at each locus. The central idea is a mathematical transformation that maps 'statistical correlation between locus pairs' to 'distance between two points in a Euclidean space'. This enables the use of geometric properties to identify proximal points (correlated locus pairs), without testing each pair explicitly. For large datasets ($\sim 10^6$ SNPs), this reduces the number of tests from 10^{12} to 10^6, significantly reducing the computational burden, without loss of power. The speed of the test allows for correction using permutation-based tests. The algorithm is encoded in a tool called RAPID (RApid Pair IDentification) for identifying paired interactions in case–control GWAS.

Results: We validated RAPID with extensive tests on simulated and real datasets. On simulated models of interaction, RAPID easily identified pairs with small marginal effects. On the benchmark disease, datasets from The Wellcome Trust Case Control Consortium, RAPID ran in about 1 CPU-hour per dataset, and identified many significant interactions. In many cases, the interacting loci were known to be important for the disease, but were not individually associated in the genome-wide scan.

Availability: http://bix.ucsd.edu/projects/rapid

Contact: vbafna@cs.ucsd.edu

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

Received on July 12, 2010; revised on September 3, 2010; accepted on September 12, 2010

1 INTRODUCTION

Recent technological developments in sequencing and genotyping have made it feasible to conduct genome-wide scans of large population cohorts to find genetic markers for common diseases (The Wellcome Trust Case Control Consortium, 2007). Nevertheless, significant challenges remain. Many genome-wide association studies (GWASs) seek to associate each marker with the disease phenotype. As multiple hypotheses are generated, individual associations must have large effect to show up as significant. In complex disorders, many independently evolving loci might interact to confer disease susceptibility, with only a modest effect at each locus. Here, we focus on detecting such interactions.

Detecting k-locus interactions in GWAS on large populations is computationally and statistically challenging, even when $k=2$. A test involving all pairs of m markers, with a case–control population of n individuals, involves $O(nm^2)$ computations. For GWAS, it is not atypical to have $n \sim 10^3$, $m \sim 10^6$ making these computations, especially with permutation-based tests of significance, intractable. A straightforward (Bonferroni-like) correction for the multiple tests would result in significant loss of sensitivity.

Therefore, many strategies for two-locus interaction testing are based on a two-stage, *filtering* approach. In the first stage (the filter stage), the objective is to discard the vast majority of locus pairs, while retaining the truly interacting pairs. If the filtering stage is fast and efficient (only a small fraction of all pairs are retained), then computationally intensive tests of association can be performed on the few remaining candidate pairs in a second, *scoring*, stage. For a filtering algorithm to be effective, it must have (a) *speed*, in that the number of computations scale linearly with the size of the data; (b) *sensitivity/power* (truly interacting pairs are retained); and, (c) *efficiency* (most pairs are discarded). Fast and efficient filters allow non-parametric permutation tests to be employed to assess significance. With the advent of deep sequencing, the number of variants considered will grow far beyond 10^6 markers, and designs of filters will be critical to GWAS analysis of interactions.

While many approaches to detecting interactions have been proposed (see Cordell, 2009, for an excellent review), the design of filters has not been investigated explicitly. A recent approach, both pragmatic and effective for filtering, is based on the assumption that interacting pairs of loci should also show a marginal effect at each locus (Marchini *et al.*, 2005). Here, the filtering stage consists of single marker tests at each locus. The scoring stage is then limited to pairs in which either one, or both loci, are individually associated. In either strategy, the filter speed is high, as single-marker analysis scales linearly with the number of loci and individuals. Empirical results show that only a small fraction of the loci show a marginal effect, leading to high efficiency. However, as Marchini *et al.* point out, there is some loss of power in employing these filters, particularly in interaction models where the marginal effects of the individual loci are small. Figure 1a provides a cartoon

*To whom correspondence should be addressed.

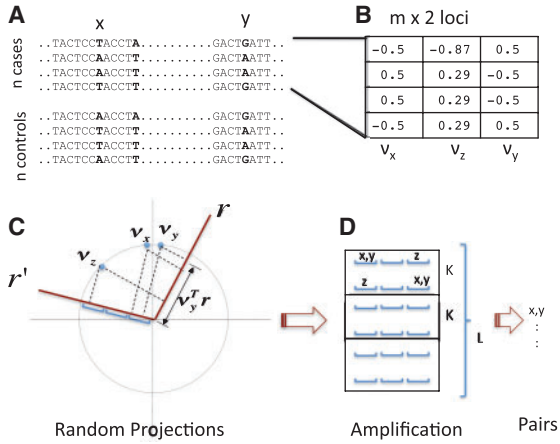

Fig. 1. Overview of RAPID. (**a**) A cartoon example of two-locus interaction with no marginal effects. Each of the variants is a coding SNP that changes the conformation of the encoded protein in a compensating manner. While neither of the two loci are individually associated, the allelic combinations A…A and T…G, mediate a disease phenotype, whereas control individuals have the combinations T…A or A…G. (**b**–**d**) An overview of the Rapid approach. Each variant locus x is transformed to a unit vector $\boldsymbol{v}_x$, and projected onto a random vector $\boldsymbol{r}$. (**c**) The length of the projection is given by $\boldsymbol{v}_x \cdot \boldsymbol{r}$. projections of interacting locus pairs x, y fall into the same bin with higher probability, relative to non-interacting pairs x, z. (**d**) Pairs that fall in the same bin K consecutive times in one of L trials are selected.

illustration of such confounding interactions. Here, compensating mutations in coding SNPs (T and G, or A and A) allow the encoded proteins to interact, but individual mutations destroy the lock and key mechanism. Therefore, the locus pair (x, y) will show strong association, but there is no marginal effect at either locus. We tackle this case.

In this article, we describe a filtering strategy, RAPID (RApid Pair IDentification). Under certain assumptions, the algorithm provides explicit guarantees on speed, efficiency and sensitivity. To formalize the argument, we parametrize the total computation for n individuals and m markers. Let input parameter ε denote the desired false negative rate of detecting interactions. RAPID performs at most $\tau_1 \simeq m^{1.07}$ tests, and allows no more than $\tau_2 \simeq m^{1.07} \ln(1/\varepsilon)$ pairs by chance, while capturing a fraction $1-\varepsilon$ of the truly interacting pairs. The surviving pairs can be tested for interaction using a total of $O(nm^{1.07} \ln(1/\varepsilon))$ computations. This can be compared with the time of $O(nm^2)$ when no filtering is employed. For GWAS, where $m \simeq 10^6$, this results in several orders of magnitude speed up. Additionally, increasing desired sensitivity $1-\varepsilon$ incurs only modest increases in running time. Extensive power simulations demonstrate the power of our approach.

We also used RAPID to reanalyze data from The Wellcome Trust Case Control Consortium (WTCCC) dataset, a benchmark GWAS. The filtering using RAPID on either dataset only took about 45 min on a 1.8 GHz, 16 GB RAM computer and identified many significant interactions.

2 THE DESIGN OF RAPID

Three key ideas underlie the design of RAPID. The first is a mapping of allelic values at loci to a Euclidean metric so that statistical correlation between two loci corresponds to distance in the Euclidean space. The second is the use of a randomized protocol called *Locality Sensitive Hashing* (LSH). LSH has been used successfully to cluster and query high-dimensional datasets (Indyk and Motwani, 1998). Its use in the context of filtering for interacting loci is novel and illustrated below. The final idea is a standard one of amplifying a probability bias through repeated trials.

Statistical correlation and the Euclidean metric: we describe a transformation that maps every locus to a point in a Euclidean Space. For exposition purposes, consider haploid data with n individuals, and m biallelic loci. Each locus x can be described by a vector $\boldsymbol{x} \in \{0,1\}^n$ of allelic values. Likewise, the case—control status of the individuals can also be expressed by a vector $\boldsymbol{d} \in \{0,1\}^n$. The null hypothesis of no association of a pair of loci x, y against $\boldsymbol{d}$ can be tested using a χ^2 test on a $2 \times 2 \times 2$ contingency table (or $3 \times 3 \times 2$ for genotype data). Specifically, the goal is to identify all locus pairs x, y such that χ^2 is high. While other, more powerful, tests of association have been proposed, the χ^2 test is sufficient for the filtering stage. By choosing appropriate thresholds, we can recover all candidate paired interactions, and leave the advanced tests of association for the scoring stage. We note that if $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{d}$ jointly associate, then at least one of the following projected associations must hold (Supplementary Section S1).

- Marginal association between $\boldsymbol{x}$ and $\boldsymbol{d}$, described by the statistic $\chi^2_{\boldsymbol{x},\boldsymbol{d}}$.
- Marginal association between $\boldsymbol{y}$ and $\boldsymbol{d}$, described by $\chi^2_{\boldsymbol{y},\boldsymbol{d}}$.
- Association between $\boldsymbol{x}$, and $\boldsymbol{y}$, when the individuals are drawn only from cases. $\chi^2_{\boldsymbol{x},\boldsymbol{y}}$ is high for cases.
- Association between $\boldsymbol{x}$, and $\boldsymbol{y}$, for control individuals. $\chi^2_{\boldsymbol{x},\boldsymbol{y}}$ is high for controls.

Each of these conditions can be tested independently. The first two describe marginal associations that can be tested directly ($O(m)$ tests). The latter two conditions have also been proposed in the context of interactions (Yang *et al.*, 1999), but require testing every pair of loci ($O(m^2)$ tests in principle). These case-only (or control-only) tests are the focus of our article. Much of the discussion below is therefore limited to a population consisting entirely of n case (or control) individuals. We use the following geometric transformations. Let P_x denote the fraction of individuals in the population with allele '1' at locus x. For $a \in \{0,1\}$, define

$$v_x(a) = \frac{a - P_x}{\sqrt{n}\sqrt{P_x(1-P_x)}}.$$

Define a vector $\boldsymbol{v}_{\boldsymbol{x}} = [v_x(x_1)\, v_x(x_2) \ldots v_x(x_n)]$. This maps the allelic values at locus x for all n individuals onto a unit vector $\boldsymbol{v}_{\boldsymbol{x}}$ (a point on the unit hypersphere) in $\Re^n$. For two loci x, y, the vectors $\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}$ are both on the unit hypersphere. Inverting the alleles at y (changing all 0's to 1's and vice-versa) does not change the correlation between the loci, but replaces $\boldsymbol{v}_{\boldsymbol{y}}$ by $-\boldsymbol{v}_{\boldsymbol{y}}$. We fold these two cases together by defining the distance between loci x, y as

$$\mathrm{dist}(\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}) = \min(\|\boldsymbol{v}_{\boldsymbol{x}} - \boldsymbol{v}_{\boldsymbol{y}}\|, \|\boldsymbol{v}_{\boldsymbol{x}} + \boldsymbol{v}_{\boldsymbol{y}}\|).$$

We can show (Supplementary Section S2) that for a pair of loci x, y,

$$\mathrm{dist}(\boldsymbol{v}_{\boldsymbol{x}}, \boldsymbol{v}_{\boldsymbol{y}}) = \sqrt{2 - 2\sqrt{\chi^2_{\boldsymbol{x},\boldsymbol{y}}/n}}. \quad (1)$$

Thus, we can transform the statistical problem of identifying interacting locus pairs $\{(x,y): \chi^2_{x,y} \geq t\}$ (where t is a threshold) into a geometric problem of computing proximal vectors: identify all locus pairs (x,y) such that

$$\text{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) \leq \theta = \sqrt{2-2\sqrt{t/n}}. \quad (2)$$

The transformations for genotypes are more complex and will be described in Section 3. By itself, the metric connection offers no help for filtering. We have m vectors in $\Re^n$, and potentially, each pair must be compared with identify the proximal pairs. Here, we apply the second idea.

LSH: to identify locus pairs (x,y) for which $\text{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) \leq \theta$, we choose a random unit vector $\boldsymbol{r}$, and project each of the points onto $\boldsymbol{r}$ (Fig. 1C). The operation is fast ($O(nm)$ steps) as each locus is treated independently. We choose an appropriate value of B (Section 3) and add each locus x to a bin numbered

$$\text{HASH}(x, \boldsymbol{r}, B) = \left\lfloor \frac{|\boldsymbol{v_x} \cdot \boldsymbol{r}|}{B} \right\rfloor$$

The bin size B is chosen to ensure that if $\text{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) < \theta$, then loci x,y fall in the same bin with high probability (denoted by f_1). On the other hand, if x,y are non-interacting ($\text{dist}(\boldsymbol{v_x}, \boldsymbol{v_y})$ is large), they fall into the same bin with a much lower probability (denoted by $f_2 < f_1$). The choices of θ and B determine possible values of f_1 and f_2. For example, we show that for $\theta = 0.1$, it is possible to construct hash functions where $f_1 = 0.95$, and at least 50% of all pairs are discarded ($f_2 \leq 0.5$). The efficiency of 50% is insufficient as we still have to test $\frac{1}{2}\binom{m}{2}$ pairs in the second stage, and have already lost 5% of the true interactions. The final idea is to amplify the bias f_1/f_2.

Amplification of bias: let $1-\varepsilon$ denote the desired power (the fraction of true interacting pairs that are retained for a second-stage scoring). Parameters L,K are computed automatically by RAPID (as described below).

We run the hashing procedure LK times, and select only those pairs that fall in the same bin all K times, in at least one of the L iterations. We will show (Section 3) that using

$$K = \frac{\ln m}{\ln(1/f_2)}, \qquad L = {f_1}^{-K} \ln(1/\varepsilon) \quad (3)$$

RAPID will output a high fraction $(1-\varepsilon)$ of all interacting pairs, but at most $m^c \ln(1/\varepsilon)$ non-interacting pairs, where $c = 1 + \frac{\ln(1/f_1)}{\ln(1/f_2)}$. We make the reasonable assumption that the number of truly interacting pairs is small. Therefore, the number of pairs output by RAPID is also approximately $m^c \ln(1/\varepsilon)$.

We show (Section 3) that the running time for RAPID is bounded by $O(nm^c \ln(1/\varepsilon))$ steps. Further, any second-stage scoring step to identify the truly interacting pairs is also bounded by $O(nm^c \ln(1/\varepsilon))$ steps ($O(n)$ steps for each pair). Substituting $f_1 = 0.95, f_2 = 0.5$, this gives a run time of $O(nm^{1.07} \ln(1/\varepsilon))$ steps for filtering and second-stage scoring. Clearly, for large m, we get a substantial speed-up in running time over the naive $O(nm^2)$ computations while maintaining the desired sensitivity.

3 MATERIALS AND METHODS

RAPID overview: the input of RAPID includes a SNP genotype matrix (limited to cases, or to controls), and two parameters θ, ε. The output is a filtered list of paired loci x,y with $\text{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) \leq \theta$ [Equation (2)], or a high value of $\chi^2_{\boldsymbol{x},\boldsymbol{y},\boldsymbol{d}}$.

The overall algorithm is described below, with details in subsequent sections. It is applied to cases only, and to controls only, and the results are merged. We map genotypes at each locus to vectors in $\Re^n$. Next, we compute the bin size B for hashing, as well as parameters L,K. Pairs that map to the same bin all K times in at least one of L iterations *and* pass an optional second stage filtering are returned. While we employ a simple second stage filtering in our tests, we expect that RAPID will be used in conjunction with sophisticated tests.

procedure GETINTERACTINGPAIRS($M_{\text{cases}}, M_{\text{controls}}, \theta, \varepsilon$)
  Set $P_1 = \text{RAPID}(M_{\text{cases}}, \theta, \varepsilon)$
  Set $P_2 = \text{RAPID}(M_{\text{controls}}, \theta, \varepsilon)$
  Output SECONDSTAGEFILTERING($P_1 \cup P_2$)

procedure RAPID(M, θ, ε)
1. Set $M' = \text{COMPUTEHAPLOIDVECTORS}(M)$
2. Set $(f_1, f_2, B, L, K) = \text{COMPUTEBINSIZE}(\theta, \varepsilon, m, n)$
3. Repeat L times
   3.1. Repeat K times
      3.1.1. Choose a unit random vector $\boldsymbol{r} \in \Re^n$
      3.1.2. For each $\boldsymbol{v_x} \in M'$
         3.1.2.1 Add $\boldsymbol{v_x}$ to the bin $\text{HASH}(\boldsymbol{v_x}, \boldsymbol{r}, B)$
   3.2. For all pairs $\boldsymbol{v_x}, \boldsymbol{v_y}$ that fall in the same bin all K times,
      3.2.1. Add (x,y) to List
4. Return List

Computing haploid vectors: The metric property of statistical correlations for the haploid case can be extended to genotypes, but the transformations are more complex. We transform the genotype vectors to haploid vectors. Consider a single locus with major allele denoted by a and minor allele denoted by A. Let n_{aa}, n_{aA}, n_{AA} denote the counts of the three genotypes. A 3×3 contingency table describes the interaction of the two locations.

A natural decomposition is to decompose the locus into $2n_{aa} + n_{aA}$ 'a' alleles, and $2n_{AA} + n_{aA}$ 'A' alleles. Note that this transformation is silent on deciding the phasing of heterozygotes for pairs. In checking interactions between two loci, the heterozygotes were phased either randomly, or identically. Neither of these choices works, as subtle interaction effects of minor variants are removed. See Supplementary Figure S2 for comparisons. Instead, we map each genotype locus x onto two haploid loci X_0, X_1. In the first, heterozygous individuals are counted as 'a' (major allele), and in the second, they are counted as 'A' (minor allele). Specifically,

$$X_0[i] = \begin{cases} 0 & \text{if } x[i] = \text{'}aa\text{'} \text{ or } x[i] = \text{'}aA\text{'} \\ 1 & \text{otherwise} \end{cases}$$

$$X_1[i] = \begin{cases} 0 & \text{if } x[i] = \text{'}aa\text{'} \\ 1 & \text{otherwise} \end{cases}$$

We show empirically (Supplementary Section S3) that for any pair x,y of loci, if x,y are interacting ($\chi^2_{\boldsymbol{x},\boldsymbol{y}}$ is high), then one of the four values $\chi^2_{X_i,Y_j}$ is high as well. The overall algorithm remains unchanged, except that all loci are transformed into biallelic loci at the beginning. At the end, a pair x,y is accepted only if at least one of its four haploid pairs X_i, Y_j is accepted.

Bin Size computation and LSH: bin size B and parameters L, K are used to balance a trade-off between speed, sensitivity and efficiency, and form the heart of the argument. For each SNP x, the procedure HASH($\boldsymbol{v_x}, \boldsymbol{r}, B$) maps $\boldsymbol{v_x}$ to

$$\left\lfloor \frac{|\boldsymbol{v_x} \cdot \boldsymbol{r}|}{B} \right\rfloor$$

Let f_1 denote the probability that two interacting pairs fall in the same bin, and let f_2 denote the probability that two non-interacting pairs fall in the same bin. We must choose bin size B so that interacting loci ($\text{dist}(\boldsymbol{v_x}, \boldsymbol{v_y}) \leq \theta$) fall in the same bin with higher probability than non-interacting pairs. Note that

$$\left\| \boldsymbol{v_x} - \boldsymbol{v_y} \right\|^2 = 1 + 1 - 2\boldsymbol{v_x} \cdot \boldsymbol{v_y} = 2 - 2\boldsymbol{v_x} \cdot \boldsymbol{v_y}$$

If x, y are non-interacting, then expected value $\mathrm{E}(\boldsymbol{v_x} \cdot \boldsymbol{v_y}) = 0$ and $\mathrm{E}(\left\| \boldsymbol{v_x} - \boldsymbol{v_y} \right\|) = \sqrt{2}$. The separation between θ and $\sqrt{2}$ allows us to choose an appropriate bin size B. Consider $z = (\boldsymbol{u} - \boldsymbol{v}) \cdot \boldsymbol{r}$ for a random vector $\boldsymbol{r} \in \Re^n$ on the unit hypersphere ($\|r\| = 1$). Observe that $\mathrm{E}(z) = 0$ and

$$\mathrm{Var}(z) = \mathrm{E}(z^2) - \mathrm{E}(z)^2 = \mathrm{E}[\sum_i (u_i - v_i)^2 r_i^2] = \frac{\|\boldsymbol{u} - \boldsymbol{v}\|^2}{n}$$

The distribution of $(\boldsymbol{u} - \boldsymbol{v}) \cdot \boldsymbol{r}$ can be approximated by $\mathcal{N}(0, \|\boldsymbol{u} - \boldsymbol{v}\| / \sqrt{n})$ (Supplementary Section S5). Let $\Phi_{\mu,\sigma}$ denote the c.d.f of a $\mathcal{N}(\mu, \sigma)$ distributed variable. If x, y are interacting, then

$$\begin{aligned} f_1 &= \Pr(\left|(\boldsymbol{v_x} - \boldsymbol{v_y}) \cdot \boldsymbol{r}\right| < B) \\ &= \Phi_{0,\theta/\sqrt{n}}(B) - \Phi_{0,\theta/\sqrt{n}}(-B) = 1 - 2\Phi_{0,\theta/\sqrt{n}}(-B) \end{aligned} \tag{4}$$

If x, y are non-interacting then

$$f_2 = \Pr(\left|(\boldsymbol{v_x} - \boldsymbol{v_y}) \cdot \boldsymbol{r}\right| < B) = 1 - 2\Phi_{0,\sqrt{2/n}}(-B). \tag{5}$$

Discrete examples of f_1, f_2, B based on different values of θ are shown in Supplementary Table S1 for illustration.

Selecting parameters L, K for two-stage LSH: the probability that two interacting pairs x, y (respectively, two non-interacting pairs) fall in the same bin each time in K random hashings is ${f_1}^K$ (respectively, ${f_2}^K$). Finally, we repeat the super-trials L times. A pair (x, y) survives if it falls in the same bin all K times, in at least one of L iterations. Therefore, a fraction $L{f_1}^K$ of the interacting pairs are selected, while only a fraction $L{f_2}^K$ of non-correlated pairs are selected. The run time of this over all m markers is $KLmn$ steps (since it takes n steps to compute each individual hash).

Various constraints are used to determine the optimum values of L, K. The probability that a truly interacting pair fails in each of the L super-trials is bounded by

$$(1 - {f_1}^K)^L \leq e^{-{f_1}^K L}$$

By choosing $L = {f_1}^{-K} \ln(1/\varepsilon)$, we ensure that $e^{-{f_1}^K L} \leq \varepsilon$, giving us the desired sensitivity of $1 - \varepsilon$. Next, we consider the total running time. The filtering time is $LKmn$ steps. Assuming that the majority of the pairs are non-interacting, the total number of pairs that survive the filtering is given by

$$\binom{m}{2} L {f_2}^K$$

Each surviving pair must be tested for interaction via a statistical test, which is typically $O(n)$ computations. Therefore, the total number of computational steps (filtering and testing) is bounded by

$$LKmn + \binom{m}{2} L {f_2}^K n \tag{6}$$

To balance the two stages, we choose $K = \frac{\ln(m)}{\ln(1/f_2)}$ so that

$$\binom{m}{2} L {f_2}^K n = \binom{m}{2} \left(\frac{f_2}{f_1}\right)^K n \ln(1/\varepsilon) \approx \frac{1}{2} m^{1 + \frac{\ln(1/f_1)}{\ln(1/f_2)}} n \ln(1/\varepsilon)$$

and

$$LKmn = \frac{Kmn \ln(1/\varepsilon)}{{f_1}^K} \approx m^{1 + \frac{\ln(1/f_1)}{\ln(1/f_2)}} \frac{\ln m}{\ln(1/f_2)} n \ln(1/\varepsilon).$$

The bin size B is now determined by choosing B that minimizes the run time (6) subject to the specified sensitivity $1 - \varepsilon$. Note that f_1, f_2 are given in terms of B by Equations (4) and (5), and K, L are determined by Equation (3). The minimization is computed numerically and occurs in the vicinity of $B \approx 2\theta/\sqrt{n}$. Altogether, L, K, f_1, f_2, B are computed in terms of inputs $n, m, \theta, \varepsilon$.

3.1 Second stage filtering

Each pair output by RAPID is considered in turn. Denote as $\chi^2_{9\times2}$, the χ^2-statistic ($df = 8$) on the 9×2 contingency table connecting the possible genotypes at a locus pair x, y. For all pairs, we compute a $\chi^2_{9\times2}$ statistic, correlating the two locus genotypes against the case–control phenotypes. Note that the $\chi^2_{9\times2}$ table gives a more relevant statistic, but the $\chi^2_{3\times3\times2}$ is easier to decompose for filtering. The $\chi^2_{9\times2}$ can be decomposed (Lancaster, 1969) as

$$\chi^2_{9\times2} = \chi^2_{\boldsymbol{x},\boldsymbol{d}} + \chi^2_{\boldsymbol{y},\boldsymbol{d}} + \chi^2_{(\boldsymbol{xy}),\boldsymbol{d}}$$

where the first two terms (each with $df = 2$) represent the marginal contributions at locus x and y, and the third term $\chi^2_{(\boldsymbol{xy}),\boldsymbol{d}}$ is approximately χ^2 distributed with $df = 4$, representing the interaction component. In second-stage filtering, we compute $\chi^2_{9\times2}$ and $\chi^2_{(\boldsymbol{xy}),\boldsymbol{d}}$ statistics. All pairs that match a user-defined cut-off on the two statistics are considered. Note that in the haplotype case, the connection between the parameter θ and the desired χ^2 cut-off is explicit, but not in the genotype case. In practice, we usually take a small value of θ. With small θ values, the most significant pairs have high sensitivity, and the sensitivity reduces for less significant interactions.

As a final step in filtering, we eliminate pairs that are too close, or are unreliable. Specifically, for the WTCCC study, the human Gap track table was downloaded from the UCSC genome browser. All SNPs within a distance 1 Mb from intervals marked in the gap table as centromeric, telomeric, heterochromatin or short arm, are discarded, and all SNP pairs on the same chromosome within a distance 1 Mb of each other are discarded. The second stage filtering is not optimal, and possibly discards truly interacting pairs. We only use these final steps to identify interesting examples of interacting pairs. We expect that RAPID filtering will be used in combination with a user-specified second stage of scoring.

3.2 Simulated datasets

To test speed versus power, we took a collection of 50K SNPs from the WTCCC dataset and exhaustively computed all pairs with $\theta \leq 0.4$, and $\theta \leq 0.1$, respectively. These served as positive controls

in testing RAPID with parameter $\theta=0.1$, and $\theta=0.4$, respectively. For different choices of ε, RAPID automatically computes parameters L,K,B. Let TP (respectively, FN) denote the number of pairs from the positive set that were filtered (respectively, discarded) by RAPID. Sensitivity was empirically computed as TP/(TP+FN). Speed-up was measured by the ratio of the number of steps in the naive computation to the number of steps in RAPID

$$\frac{nm^2}{LKmn+\binom{m}{2}Lf_2{}^K n}$$

To test speed versus data-size, we duplicated the WTCCC SNPs and appropriate number of times to create datasets of size up to $100M$ SNPs. Both speed-up, and actual runtimes were plotted against the threshold θ, after fixing $\varepsilon=0.05$.

To compare RAPID against other filtering approaches, we simulated two datasets according to two general epistatic models that have been proposed earlier (Marchini *et al.*, 2005). See Supplementary Section S4 for details. In Model 1, the odds of disease increase multiplicatively with genotype both within and between loci. Model 2 (corresponding to Model 3 from Marchini *et al.*), has threshold effects so that the odds of disease increase only if both loci have at least one disease-associated allele, but additional minor alleles do not further increase the odds. Marginal effects for each locus (independent of the other) should be present in Model 1, but less so for Model 2. For each model, we simulated genotypes at two interacting loci in $n=2000$ cases and $n=2000$ controls. The distribution of genotypes was governed by the following parameters: marginal odds for major alleles and minor allele-specific effects described by parameter β. Empirical values of these parameters are difficult to obtain. Instead, we followed the calculations of Marchini *et al.* to derive them using λ, where $1+\lambda$ is the heterozygous odds ratio at a single locus, disease prevalence $P_D=0.01$, and minor allele frequencies $\pi=\pi_A=\pi_B$.

For each choice of parameters $\lambda\in\{0.2,0.5,1.0\}$ and $\pi\in\{0.05,0.1,0.2,0.5\}$, and derived values of α and β, we simulated 1000 interacting pairs, by sampling genotypes values according to probabilities induced by the model parameters. Significance of search results was corrected using $m=300\,000$ SNPs.

3.3 WTCCC case–control dataset

We tested the performance of RAPID on a published dataset: the WTCCC consortium dataset (~2000 individuals in six diseases [Hypertension (HT), Bipolar Disorder (BD), Crohn's disease (CD), Coronary Artery Disease (CAD) and Type 1 and 2 Diabetes (T1D, T2D)], against ~3000 shared controls, genotyped at ~500 000 loci. A total of 469 557 SNPs and individuals that passed the WTCCC quality controls were used.

4 RESULTS

RAPID has two key parameters, θ and ε. The user decides the minimum strength of interactions (measured by the χ^2 statistic), and uses it to specify a bound on the Euclidean distance θ between the transformed vectors [Equation (1)]. The parameter ε describes the acceptable type II error rate among the pairs that satisfy the θ threshold. The parameters L,K also influence speed and sensitivity, but can be computed given n,m,θ,ε, as described in Equation (3).

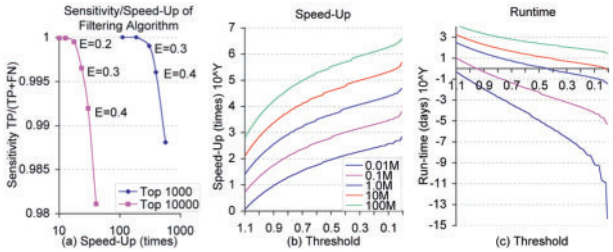

Fig. 2. Speed Speed sensitivity trade-offs in RAPID. The trade-offs are computed as a function of user-defined parameters θ,ε. (**a**) Speed-up versus sensitivity trade-offs are measured on a dataset of 50 000 WTCCC control SNPs. Different thresholds θ are chosen to filter the top 1000 ($\theta\leq0.1$), 10 000 ($\theta\leq0.4$) of SNP pairs, respectively. $\varepsilon\in\{0.05,0.1,0.2,0.3,0.4,0.5\}$. (**b** and **c**) Runtime versus θ. Strongly interacting pairs ($\theta\leq0.5$) can be identified on large datasets ($1M$ SNPs, 3000 genotypes) with 95% sensitivity in a few hours on a commodity PC. All experiments were run on a 1.8 GHz, 16 GB RAM, Linux machine. Axes have logarithmic scale.

We tested the speed versus sensitivity of RAPID on the WTCCC control dataset of 3000 individuals, using a different choice of θ,ε. For each choice of θ, define the false negatives (FN) as the number of pairs that exceed the threshold θ, but were rejected by RAPID. True positives (TP) is the number of pairs exceeding the threshold that were accepted. Likewise, for each choice of θ,ε, we define speed up as the drop in number of computations, compared with the naive approach ($O(nm^2)$ computations). Figure 2a describes the speed versus sensitivity [TP/(TP + FN)] trade-off for $\theta=0.1$ and 0.4. For low values of θ, we can show 2–3 orders of magnitude speed up without sacrificing sensitivity. We next explored the speed-up, and actual runtimes for increasingly large datasets (Fig. 2b–c), and different thresholds after fixing the $\varepsilon=0.05$ (or 5%). Even for large datasets ($1M$ SNPs and 3000 genotypes), strongly interacting pairs can be identified with 95% sensitivity in a few hours on a commodity PC. In practice, we work with low θ values for large datasets as only the most significant interactions are required.

We compared RAPID against two strategies widely used to measure interactions (Evans *et al.*, 2006; Marchini *et al.*, 2005). Strategy I can be thought of as a filtering strategy. Only a subset of the possible pairs are tested, those in which one of the SNPs (Strategy Ia), or both (Strategy Ib) show a marginal effect. In Strategy II, all pairs of SNPs are tested for interaction. On the face of it, Strategy II should demonstrate the highest power. However, Evans *et al.* argue that as the scoring stage is limited to the filtered pairs, the correction factor is based on the number of pairs being considered at the scoring stage. To ensure an overall type I error of at most α, the significance level of the test for each pair is set to α/τ_2 where τ_2 is the number of filtered pairs. For Strategy I, $\tau_2\simeq m$, but for the full interaction model, $\tau_2=\binom{m}{2}$. Due to the larger number of tests in Strategy II, Strategy I often outperforms Strategy II (Fig. 3).

As RAPID is only a filtering strategy, it cannot be directly compared. Here, we took only the top $\tau_2=\binom{m}{2}Lf_2{}^K$ pairs output by RAPID, and used a second-stage $\chi^2_{9\times2}$ P-value with $df=8$ (Section 3) to compute Type I error α. We set the significance level for each test to $\alpha/(\binom{m}{2}Lf_2{}^K)$. We note that with the speed-ups obtained, we can also use non-parametric, permutation tests of significance.

We tested the approach on two different models of disease interaction (Section 3). In Model 1, the odds of disease increase

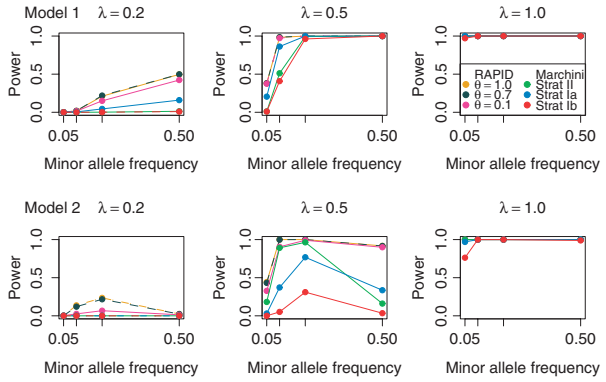

Fig. 3. Power of RAPID. Power comparison of strategies used in GWASs. Interacting loci are simulated (300K SNPs with 2000 cases and 2000 controls) according to two general epistatic models, first described by Marchini *et al.* In the first model, the case odds increase multiplicatively upon addition of minor alleles at either loci. In the second (threshold) model, the odds increase only if both locus have at least one minor allele, and do not improve with addition of minor alleles. The model parameters in both cases are derived numerically using the single locus heterozygous odds ratio $1+\lambda$ ($\lambda \in \{0.2, 0.5, 1.0\}$), and disease incidence $P_D = 0.01$. Model 1 (top panels) shows marginal effects, as do models with higher values of λ ($=1$). RAPID outperforms other methods, and shows a clear improvement when the marginal effects are low.

multiplicatively with genotype both within and between loci. Model 2 (corresponding to Model 3 from Marchini *et al.*), has threshold effects so that the odds of disease increase only if both loci have at least one disease-associated allele, but additional minor alleles do not further increase the odds. For each model, we simulated genotypes at two interacting loci in $n = 2000$ cases and $n = 2000$ controls. The distribution of genotypes was governed by the parameter λ, where $1+\lambda$ is the heterozygote odds ratio. For each choice of parameters, we simulated 1000 pairs, and corrected using $m = 300000$ SNPs. Figure 3 shows power of each strategy for different simulation parameters. Low values of λ present the higher challenge, as association exists with low marginal effects.

In all tests, the performance of RAPID shows greatly improved power over previous approaches. Our analysis reveals that many significant pairs do not show individual effects and might be missed by Strategy I, but are identified by RAPID without the overhead of multiple tests.

4.1 Results on the WTCCC data

We analyzed six WTCCC datasets with RAPID, using stringent parameters $\theta = 0.1$, $\varepsilon = 0.05$ for speed, and to detect the most significant interactions first. The search time for each data set was about 75 min, with a total of $\sim 10M$ pairs output prior to second-stage filtering. For second-stage filtering, we filtered out proximal pairs (Section 3), and only kept pairs whose $\chi^2_{9\times 2}$ P-value was 10^{-7} or better, and there was a significant interaction component (the $\chi^2_{(xy),d}$ $P \leq 0.01$). This resulted in a total of 34 hits across all datasets (Supplementary Table S2). A select subset of pairs with a stronger interaction component is shown in Table 1.

BD: we find a single pair [rs13126272 (chr4), rs1918942 (chr12)] with interaction above the threshold—the $\chi^2_{9\times 2}$ P-value is 8.9×10^{-22}. The SNP rs13126272 is proximal to the solute carrier protein ANT1. Mutations in the ANT1 gene mediate the BD phenotype (Siciliano *et al.*, 2003). The SNP rs1918942 is proximal to contactin (CNTN1), a neural adhesion gene that is implicated in neurological disorders including BD (O'Dushlaine *et al.*, 2010). In the WTCCC dataset, the two SNPs show a strong interaction (P-value 6.3×10^{-3}).

Table 1. Significant gene–gene interactions in the WTCCC dataset

SNP A	SNP B	$\chi^2_{9\times 2}$ P-value	$\chi^2_{(xy),d}$ P-value
BD			
rs13126272 (chr4)	rs1918942 (chr12)	8.88e−22	6.28e−3
CAD			
rs4970605 (chr1)	rs11216700 (chr11)	1.42e−17	6.90e−3
rs12061996 (chr1)	rs2658728 (chr12)	1.07e−8	3.79e−3
rs2369810 (chr1)	rs2164411 (chr2)	4.35e−13	6.61e−3
rs2164411 (chr2)	rs1684835 (chr5)	1.27e−12	7.55e−3
rs2164411 (chr2)	rs17139253 (chr5)	6.97e−14	9.31e−4
rs2164411 (chr2)	rs7840975 (chr8)	2.04e−13	4.11e−3
rs2164411 (chr2)	rs17369334 (chr10)	4.06e−13	7.11e−3
rs2164411 (chr2)	rs906467 (chr20)	1.25e−12	6.10e−3
rs9875049 (chr3)	rs2658728 (chr12)	2.35e−9	8.11e−3
rs1828416 (chr4)	rs2658728 (chr12)	2.40e−9	3.32e−3
HT			
rs17146413 (chr11)	rs12590471 (chr14)	4.52e−9	5.89e−5
T1D			
rs7576174 (chr2)	rs12661352 (chr6)	2.83e−15	3.22e−3
rs6546693 (chr2)	rs3130564 (chr6)	2.46e−19	4.05e−3
rs935497 (chr3)	rs9268858 (chr6)	2.14e−41	6.62e−3
rs244545 (chr5)	rs2596571 (chr6)	2.77e−34	6.00e−3
rs6888673 (chr5)	rs2523742 (chr6)	1.89e−11	3.53e−3
rs12661352 (chr6)	rs9320240 (chr6)	8.66e−15	8.68e−3
rs12661352 (chr6)	rs4359308 (chr13)	1.72e−15	8.05e−4
rs12661352 (chr6)	rs17413237 (chr13)	5.42e−14	6.31e−3
rs12661352 (chr6)	rs17263755 (chr18)	1.23e−14	6.18e−3
rs12661352 (chr6)	rs4818677 (chr21)	7.18e−16	1.68e−3
rs12661352 (chr6)	rs6518154 (chr21)	2.76e−15	3.64e−3
T2D			
rs2164411 (chr2)	rs884289 (chr10)	6.67e−13	5.71e−3

CAD: among the interactions observed, many involve SNP rs2164411 (chr 2) in the gene DNMT3a, involved in DNA methylation. Altered global DNA methylation has been observed in CAD and other cardiovascular diseases (Movassagh *et al.*, 2010; Sharma *et al.*, 2008).

T1D: we observe a number of interactions connected to T1D. Many variants, including rs12661352 (HLA-DQB2), rs9268858 (HLA-DRA), rs2523742, rs244545 (HLA-B) involve genes on the 6p21 HLA locus which is a known susceptibility locus for T1D (Undlien and Thorsby, 2001). Consider the interaction between rs935497 and rs9268858. rs934597 lies in the PPM1L gene which has been implicated in diabetes in mouse knockouts (Chen *et al.*, 2008). It adds a significant interaction term to rs9268858 (χ^2 P-value 6.6×10^{-3}), which lies in the HLA-DR region. Likewise, consider

the interaction of rs244545 and rs2596571. The variant rs244545 lies in a gene desert, but upstream of the TF islet-1. Islet-1 is known to activate insulin gene expression, and is also implicated in pancreatic islet cell development. Mutations in ISL-1 are associated with maturity-onset diabetes of the young (Barat-Houari *et al.*, 2002).

HT: HT is a complex disorder with high heritability, but few loci have been confirmed in GWASs. The SNP rs12590471 associates weakly (P-value 6.08×10^{-6}). However, the $\chi^2_{9\times 2}$ statistic with SNP rs17146413 has P-value 4.52×10^{-9}. The SNP rs17146413 lies in EHD1, involved in endocytosis of IGF-1 receptor (Rotem-Yehudar *et al.*, 2001). Circulating IGF-1 levels are known to mediate HT (Galderisi *et al.*, 2001).

5 DISCUSSION

The chief contribution of RAPID is a methodological one. We make a dent in the seemingly intractable problem of identifying pairs of loci that interactively associate with the phenotype, while not showing any marginal effects. On the one hand, the algorithm is naive, especially from a statistician's perspective. It only uses χ^2-like tests, when many, more sophisticated tests are possible. This is done with a purpose. First, RAPID is designed to be complementary to the array of tools being developed for interactions. It is meant to be used as a filtering tool, to identify a small number of candidate pairs on which any statistic can be applied. Indeed, there are many different models by which loci can interact. RAPID starts with the lowest common denominator, on the idea that any form of interactions implies a departure from the χ^2-values. The combination of RAPID with multiple second-stage scoring strategies, can help capture different interactions.

Second, our method exposes an interesting relationship between χ^2 and the Euclidean metric, which has not (to our knowledge) been previously explored. This geometrization of statistical correlations should open new algorithmic approaches for identifying multiple-locus interactions. The extension of the metric properties to the genotype case is significantly harder than haplotypes, and raises open questions about multi-allelic loci. In some of our own work, we are considering more accurate projections of the genotypes, and have extended it to multi-allelic loci as well (Tesler,G., submitted for publication).

Here, we only handle the pairwise interaction case. While we can work out a formal technique for multiple loci, there is a pragmatic way in which we can simply consider chains (or, dense, connected components of interacting locus pairs). The results on the WTCCC dataset, while interesting in themselves, are only presented as an exemplar of how RAPID can be used. We plan to collaborate with other groups and domain experts on specific association studies.

We conclude with a remark on the advent of inexpensive next generation sequencing. Clearly, we will see an explosion in the number of sites m (including rare variations), as well as the sizes of the cohorts n. Empirical false discovery rate (FDR) calculations help reduce the multiple testing problem. However, the application of these tests demands computational efficiency, providing additional motivation for algorithmic developments in genetics.

Funding: National Science Foundation IGERT training grant DGE-0504645 (to M.S.); NSF grant DMS-0718810 (to G.T.); NSF grant IIS-081090 (to V.B.).

Conflict of Interest: none declared.

REFERENCES

Barat-Houari,M. *et al.* (2002) Positional candidate gene analysis of Lim domain homeobox gene (Isl-1) on chromosome 5q11-q13 in a French morbidly obese population suggests indication for association with type 2 diabetes. *Diabetes*, **51**, 1640–1643.

Chen,Y. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.

Cordell,H.J. (2009) Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Evans,D. *et al.* (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet.*, **2**, e157.

Galderisi,M. *et al.* (2001) Inverse association between free insulin-like growth factor-1 and isovolumic relaxation in arterial systemic hypertension. *Hypertension*, **38**, 840–845.

Indyk,P. and Motwani,R. (1998) Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. ACM, New York, NY, USA, pp. 604–613.

Lancaster,H.O. (1969) *The Chi-squared Distribution*. John Wiley & Sons, Inc., New York, NY, pp. 260–270.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

Movassagh,M. *et al.* (2010) Differential DNA methylation correlates with differential expression of angiogenic factors in human heart failure. *PLoS ONE*, **5**, e8564.

O'Dushlaine,C. *et al.* (2010) Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol. Psychiatry* [Epub ahead of print; doi:10.1038/mp.2010.7; 16 February 2010].

Rotem-Yehudar,R. *et al.* (2001) Association of insulin-like growth factor 1 receptor with EHD1 and SNAP29. *J. Biol. Chem.*, **276**, 33054–33060.

Sharma,P. *et al.* (2008) Detection of altered global DNA methylation in coronary artery disease patients. *DNA Cell Biol.*, **27**, 357–365.

Siciliano,G. *et al.* (2003) Autosomal dominant external ophthalmoplegia and bipolar affective disorder associated with a mutation in the ANT1 gene. *Neuromuscul. Disord.*, **13**, 162–165.

The 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.

Undlien,D.E. and Thorsby,E. (2001) HLA associations in type 1 diabetes: merging genetics and immunology. *Trends Immunol.*, **22**, 467–469.

Yang,Q. *et al.* (1999) Case-only design to measure gene-gene interaction. *Epidemiology*, **10**, 167–170.