BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 3 2014, pages 384–391
doi:10.1093/bioinformatics/btt697

Genetics and population analysis

Advance Access publication November 28, 2013

ABACUS: an entropy-based cumulative bivariate statistic robust to rare variants and different direction of genotype effect

Barbara Di Camillo*, Francesco Sambo, Gianna Toffolo and Claudio Cobelli

Department of Information Engineering, University of Padova, via Gradenigo 6B, 35131 Padova, Italy

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: In the past years, both sequencing and microarray have been widely used to search for relations between genetic variations and predisposition to complex pathologies such as diabetes or neurological disorders. These studies, however, have been able to explain only a small fraction of disease heritability, possibly because complex pathologies cannot be referred to few dysfunctional genes, but are rather heterogeneous and multicausal, as a result of a combination of rare and common variants possibly impairing multiple regulatory pathways. Rare variants, though, are difficult to detect, especially when the effects of causal variants are in different directions, i.e. with protective and detrimental effects.

Results: Here, we propose ABACUS, an Algorithm based on a BivAriate CUmulative Statistic to identify single nucleotide polymorphisms (SNPs) significantly associated with a disease within predefined sets of SNPs such as pathways or genomic regions. ABACUS is robust to the concurrent presence of SNPs with protective and detrimental effects and of common and rare variants; moreover, it is powerful even when few SNPs in the SNP-set are associated with the phenotype. We assessed ABACUS performance on simulated and real data and compared it with three state-of-the-art methods. When ABACUS was applied to type 1 and 2 diabetes data, besides observing a wide overlap with already known associations, we found a number of biologically sound pathways, which might shed light on diabetes mechanism and etiology.

Availability and implementation: ABACUS is available at http://www.dei.unipd.it/~dicamill/pagine/Software.html.

Contact: barbara.dicamillo@dei.unipd.it

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

Received on July 4, 2013; revised on October 29, 2013; accepted on November 25, 2013

1 INTRODUCTION

In the past few years, the hereditary component of complex multifactorial diseases has started to be explored through the novel paradigm of genome-wide association studies (GWASs). A GWAS searches for patterns of genetic variation, in the form of single nucleotide polymorphisms (SNPs), between a population of affected individuals (cases) and a healthy (control) population. Although these studies have successfully identified a number of significant SNP–disease associations, they were able to explain only a small fraction of disease heritability (Manolio *et al.*, 2009). One of the reasons for this lack of success, as already faced in microarray data analysis (Di Camillo *et al.*, 2012; Sanavia *et al.*, 2012), is that complex pathologies, such as cancer, diabetes or neurological disorders, are heterogeneous and multicausal, as a result of the alteration of multiple regulatory pathways and of the interplay between different genes and the environment, rather than imputable to a single dysfunctional gene like monogenic diseases (Moore *et al.*, 2010). Another important reason is that a combination of rare and common variants is likely to contribute to the disease (Gibson, 2012). Rare variants, though, are more difficult to detect than common variants (Asimit and Zeggini, 2010); in fact, single-marker tests are not powerful enough when applied in a context of low evidence of association (relatively low number of subjects carrying the rare allele) together with the need of correction for multiple testing (Dudoit *et al.*, 2003).

Several alternatives to single marker tests have been proposed in the literature to detect rare variants. Multiple marker methods test the association of a group of variants, e.g. SNPs within the same gene or pathway, to the disease. In this context, a widely applied approach is to test for the significance of accumulation of rare alleles within a phenotype, across a group of SNPs. Briefly, for each subject, SNPs in the same group are collapsed to an indicator variable summarizing either the proportion of rare variants that carry at least one minor allele or the presence/absence of at least one rare variant (Li and Leal, 2008; Madsen and Browing, 2009; Morgenthaler and Thilly, 2007). These approaches, known as collapsing methods or burden tests, loose power when a portion of SNPs increases the risk of disease and the remaining portion is protective. Multi-marker approaches alternative to burden tests include the Hotelling two samples T^2 test (Fan and Knapp, 2003), the Zglobal statistic (Schaid *et al.*, 2005) and the weighted score test proposed by Wang and Elston (2007). The Hotelling two samples T^2 test is a generalization of the Student's *t*-test, whose degrees of freedom increase with the number of SNPs being simultaneously tested, thus losing power with the SNP-set size. The Zglobal statistic and the weighted score test have only one degree of freedom; however, using these methods implies to know the risk allele at each variant and, as for burden tests, power is affected by the relative proportions of SNPs increasing and decreasing the risk of disease.

Alternatives, whose power does not depend on the SNP-set size and that do not make any assumption on the direction of the SNP effect (i.e. on the protective or detrimental effect of the variants on the phenotype), are the methods based on genotype similarity between individual in the same group, such as

*To whom correspondence should be addressed.

multivariate distance matrix regression (Wessel *et al*., 2006) and kernel-based association tests (KBAT) (Mukhopadhyay *et al*., 2010). In particular, multivariate distance matrix regression has been shown to be more powerful for sets of correlated SNPs, whereas KBAT well handles both correlated and uncorrelated SNPs (Morris and Zeggini, 2010). In general, these latter methods based on genotype similarity between individuals are robust to the direction of genotype risk. However, they have been reported to lose power when higher minor allele frequency (MAF) SNPs are included in the SNP-set or when many SNPs are jointly analyzed and only few of them are associated with the disease, probably due to the relative low number of subjects compared with the number of possible combinations of rare variants (Asimit and Zeggini, 2010; Zeggini and Asimit, 2011). Recently, an optimal unified approach for rare variant association testing has been proposed by Lee *et al*. (2012); the method, called SKAT-O, combines burden tests with a sequence kernel association test.

Because in general, a combination of rare and common variants influencing the genotype with a protective or detrimental effect is likely to contribute to the disease, an ideal method should be robust to different MAF, to different direction of genotype effects and to the number of associated SNPs within the SNP-set being analyzed. Moreover, it is desirable to gain some knowledge on the specific SNPs associated with the disease. In fact, multi-marker and cumulative test methods assess the association of a group of markers, but do not distinguish between associated and not associated markers within the group.

Here, we propose ABACUS, an Algorithm based on a BivAriate Cumulative Statistic designed to analyze SNPs with different MAF in the same group, independently on the protective or causative effect of the minor frequency allele.

Being based on a bivariate statistic, ABACUS, differently from other methods, performs multiple tests on each SNP, namely, equal to the number of SNPs in the SNP-set minus 1. This provides multiple evidence of associations and allows increasing the sensitivity with respect to other methods, such as methods based on genotype similarity between individuals in the same group and sequence kernel association tests, which instead calculate a cumulative SNP-set statistic to associate SNP-sets to phenotype. Moreover, the bivariate statistic used by ABACUS is independent on the minor allele being protective or causative to the disease, which makes ABACUS advantageous with respect to burden tests. Relying just on a bivariate statistic, though, would be type 1 error prone because of the number of performed tests. Thus, to control type I error, ABACUS implements a Bonferroni correction within each SNP-set together with a graph-theoretic approach to identify groups of significant SNPs.

Applied to a whole SNP dataset, ABACUS gives as output a list of SNP-sets associated with the disease and, for each SNP-set, the list of significant SNPs.

ABACUS, like other methods, first requires the definition of the SNP-sets, such as pathways, genes or genomic regions encoding a priori information on the potential point effects of the SNPs in each subset. We consider biological pathways as the preferred definition of SNP-sets, as studying the cumulative variation of SNPs mapping on genes in the same pathway (interacting genes) might fill in part the missing heritability and guide mechanistic studies helping uncovering the underlying disease pathways (Barrett *et al*., 2009). Moreover, ABACUS is particularly suited for pathway analysis, given its ability of simultaneously considering common and rare variants and different direction of genotype effects.

In the following, we introduce ABACUS and assess its performance in comparison with other methods on a number of simulated datasets with known genotype–phenotype associations. To better appreciate the various facets of the method and assess it also on real data, we illustrate ABACUS application to the Wellcome Trust Case Control Study on type 1 and 2 diabetes (The Wellcome Trust Case Control Consortium, 2007). Results show how ABACUS is able to select a number of SNPs and genes associated to diabetes, previously identified either by the WTCCC consortium or in different T1D and T2D GWAS, showing high sensitivity in detecting both common and rare variants. Moreover, ABACUS identifies new biologically sound associations with diabetes, involving genes associated with focal adhesion, platelet homeostasis, inositol phosphate metabolism and glutathione metabolism for type 1 diabetes and tryptophan metabolism and lipid homeostasis for type 2 diabetes.

2 METHODS

ABACUS exploits (i) a bivariate statistic, from now on called S_2, calculated for each pair of SNPs within the SNP-set and (ii) an aggregated score measuring the cumulative evidence of genetic–phenotypic association of the SNPs annotated in the SNP-set. We first introduce the statistic S_2 and then describe the ABACUS algorithm.

2.1 Computing the statistic S_2 for a pair of SNPs

The genotype of a pair of SNPs i and j with alleles A,a and B,b, respectively, has nine possible configurations: AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb and aabb. The main building block of a SNP pair-based analysis is thus a 2×9 contingency table, reporting the frequencies of the number of genotypes in the case and control conditions. In Table 1, f_{cg} are the frequencies of the number of cases ($c=1$) and controls ($c=2$) with genotype g ($g=1,..,9$), i.e. the counts of the different genotypes, divided by the number of samples in cases (N_1) and controls (N_2), respectively. Thus, in the hypothesis of no association between the pair of SNPs and the disease, $f_{1g} \cong f_{2g}$ ($\forall g=1,..,9$).

To test this hypothesis we exploit the concept of entropy (Shannon and Weaver, 1963) and define a statistic S_2, obtained by calculating, for each genotype g, the relative weighted difference between the expected and the observed entropy and summing it across the nine genotypes:

$$S_2(i,j) = \sum_{g=1}^{9} \left(\frac{H_0 - H_g}{H_0} \right) \cdot F_g \qquad (1)$$

where: F_g is the proportion of the genotype g in the entire dataset, i.e. $F_g = (f_{1g} \cdot N_1 + f_{2g} \cdot N_2)/(N_1 + N_2)$; H_g is the entropy of genotype g in the two populations of cases and controls, computed as follows:

$$H_g = -\frac{f_{1g}}{F_g} \cdot \log_2\left(\frac{f_{1g}}{F_g}\right) - \frac{f_{2g}}{F_g} \cdot \log_2\left(\frac{f_{2g}}{F_g}\right) \qquad (2)$$

and H_0 is the maximum entropy value, occurring when $f_{1g} = f_{2g}$; i.e. in a two classes problem $H_0 = 1$. Genotypes with frequency 0 have $H_g = 0$, according to the entropy definition.

It has to be noted that a small difference, e.g. of one single subject, in the number of cases with genotype g with respect to the controls implies either a large or a negligible difference between f_{1g} and f_{2g} depending on g being rare or common, respectively. As a consequence, H_g will be much

Table 1. Contingency table for case/control joint analysis of a pair of SNPs

Group/ Genotype	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb	Σ
Cases	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	$\Sigma f_{1g}=1$
Controls	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	f_{27}	f_{28}	f_{29}	$\Sigma f_{2g}=1$

lower than H_0 in the first case, thus contributing to increase $S_2(i, j)$, and almost equal to H_0 in the second, thus giving almost no contribution to $S_2(i, j)$. In general, this is a desirable property for H_g and, with a number of subjects tending to the whole population, unbiased. However, in practice, the number of subjects is limited to the observed sample; thus rare genotypes counts, limited to some units, can give an inaccurate estimate of f_{1g} and f_{2g}. In Equation 1, the relative difference between the observed entropy H_g and its maximum value H_0 is thus weighted by F_g, to correct for possible biases introduced by genotypes with different frequencies.

The higher is S_2, the higher is the confidence that the two SNPs are associated with the phenotype. The distribution of S_2 under the null hypothesis is obtained by repeatedly shuffling at random subject labels from cases and controls and considering the distribution of all test statistics, i.e. for all the SNP pairs (i, j).

A statistic S_1 analogous to S_2 can also be defined for a single SNP i, using a 2×3 contingency table. $S_1(i)$ represents the strength of the association between locus i and the phenotype; the difference $\Delta S_2(i, j)$ between $S_2(i, j)$ and max $\{S_1(i), S_1(j)\}$ can thus be interpreted as the fraction of association between genotype and phenotype that is explained by the joint effect of the two SNPs and that cannot be explained by a single gene model. In other words, the value of the statistic $S_2(i, j)$ is enhanced by the strength of association of the two single SNPs i and j and by their possible epistatic interaction.

2.2 The ABACUS algorithm

Given a number of SNPs annotated in different SNP-sets, with possible overlaps, ABACUS analyzes each SNP-set independently and, for each SNP-set, selects the best subset of SNPs jointly associated with the phenotype. A SNP-set is considered significantly associated with the phenotype if it contains at least one SNP significantly associated with it.

In details, given P SNPs belonging to a given SNP-set, ABACUS selects significantly associated SNPs based on the following steps (Fig. 1 illustrates the algorithm with a graphical example).

(1) $S_2(i, j)$ is computed for each pair of SNPs (i, j) in the SNP-set

(2) For each SNP i, $MS(i) = \text{median}\ \{S_2(i, j); j = 1, .., P\}$ is computed.

(3) The confidence thresholds θ for $S_2(i, j)$ are fixed corresponding to a significance level α (corrected for the number of tests $P \cdot (P\text{-}1)/2$ using Bonferroni correction).

(4) An undirected graph $G = (V, E)$ with vertices $V = \{1, .., P\}$ and edges $E = \{(i, j);\ i, j \in V \mid S_2(i, j) > \theta\}$ is defined (Fig. 1b).

(5) While the edge set $E \neq \emptyset$:

- The SNP i_w with the highest $MS(i)$ is picked and, if there is at least one edge $(i_w, j) \in E$, i_w is considered associated with the phenotype, with score $MS(i)$ (Fig. 1c and e).
- All edges (i_w, j) are removed from E (Fig. 1d and f).

The first step of the algorithm is to compute $S_2(i, j)$ for each pair of SNPs in the gene set. The higher the risk carried by i and/or j, the higher $S_2(i, j)$ is expected to be. In general, if i is not associated with the phenotype, $S_2(i, j)$ will be above the threshold θ only if j is associated with the phenotype or in case of a false positive. On the other hand, if i is associated with the phenotype, $S_2(i, j)$ is likely to be above θ, with increased value in case also j is associated, with single or joint effect on the phenotype. However, relying just on $S_2(i, j)$ to directly measure the association between the pair (i, j) and the phenotype is type 1 error prone, both because every SNP i is tested $P \cdot (P\text{-}1)/2$ times and because $S_2(i, j)$ is likely to pass the threshold θ even if just one of the two SNPs (i or j) is associated with the phenotype. For these reasons, ABACUS implements a Bonferroni correction within each SNP-set (α is corrected for the number of tests $P \cdot (P\text{-}1)/2$) together with an iterative graph-pruning strategy to identify groups of significant SNPs. More in details, ABACUS ranks the SNPs within a SNP-set based on the median value $MS(i)$ of the statistic $S_2(i, j)$ across the values observed for each pair (i, j), $j = 1, \ldots, i\text{-}1, i+1, \ldots, P$, rather than on the statistic itself. Because the median value $MS(i)$ of the statistic $S_2(i, j)$ is based on multiple evidences of association of SNP i with the phenotype, it is a precise and sensitive statistic to select significant SNPs, as shown in Section 3. The median was preferred with respect to the average or maximum value based on empirical observation of the results on a number of different simulations (data not shown).

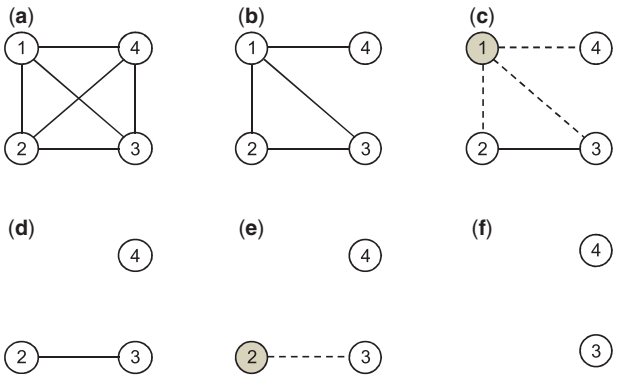

Fig. 1. Example of ABACUS working on a simple set of four SNPs (**a**) Definition of a fully connected undirected graph $G = (V,E)$; (**b**) only edges (i, j) with $S_2(i, j) > \theta$ are kept in the graph; (**c**) the SNP with highest $MS(i)$ is picked (in gray in the figure) and, if it has at least one incident edge (dashed lines in the figure), it is considered associated with the phenotype; (**d**) then its edges are removed from the graph; (**e** and **f**) iteration of step 5 on SNP 2

A null hypothesis can be derived also for $MS(i)$ and, in principle, the algorithm could stop at step 2. The rationale of steps 3–5 of the algorithm is to further improve ABACUS precision limiting the loss of sensitivity, by separating the confounding effect that strongly associated SNPs might exert on the value $MS(\cdot)$ of non-associated SNPs.

For example, referring to Figure 1, if SNP 1 is associated with the phenotype but 4 is not, $S_2(1, 4)$ will likely be above the threshold θ and $MS(1)$ will likely be greater than $MS(4)$, thus 1 will be analyzed before 4, associated with the phenotype and removed from V (Fig. 1c and d). When 4 will then be analyzed, the edge (1, 4) will have already been removed from E and so all the edges (4, j), with $MS(j) > MS(4)$ (Fig. 1f). Node 2, on the other hand, if analyzed after node 1 and 4, has another incident edge after the removal of node 1, so it will be selected.

ABACUS outputs the list of SNP-sets associated with the phenotype and, for each SNP-set, the list of SNPs i_w associated with the phonotype and their score $MS(i_w)$, which can be used to rank the SNPs in the SNP-set based on the strength of confidence of association with the disease. Because ABACUS implements a Bonferroni correction within each SNP-set followed by the graph pruning strategy described earlier in the text, the familywise error rate, i.e. the probability of making at least one false SNP-phenotype association within the SNP-set, is controlled at a

probability equal or lower than the chosen significance level α (see Section 3). Thus, when different SNP-sets (let us suppose statistically independent for the sake of simplicity) are analyzed, we expect a number of false positive SNP-sets equal to α multiplied by the number of SNP-sets that are truly non-associated with the phenotype (at maximum the number of analyzed SNP-sets). For example, if $\alpha = 0.05$ and 100 SNP-sets are analyzed, we expect a number of false positive associations ≤ 5. ABACUS does not correct automatically by the number of SNP-sets being analyzed but leaves the user to set the appropriate significance level α.

Steps 1–5 of the algorithm have computational complexity $O(N \cdot P^2)$, $O(P^2 \cdot \log P)$, $O(1)$, $O(P^2)$, $O(P \cdot \log P)$, respectively; thus, analyzing one SNP-set has computational complexity $O(P^2 \cdot (N + \log P))$, with N being the number of subjects and P the number of SNPs.

ABACUS software is built as an R package with the most computationally demanding functions written in C; it is released under the GNU general public license and is available at http://www.dei.unipd.it/~dicamill/pagine/Software.html.

2.3 Type I error simulations

To investigate whether ABACUS attains to the desired type I error rate at low significance levels, e.g. $\alpha = 10^{-6}$, it is necessary to conduct simulations with hundreds of millions of simulated datasets. To do that while diminishing the computational burden produced by simulating and analyzing such an amount of data, we generated 10 datasets of 4000 subjects and 200 000 SNPs and, for each dataset, we repeatedly (100 times) sampled 10 000 SNP-sets of 20 SNPs each, for a total of 10^7 simulated SNP-sets. For each SNP, the MAF was randomly sampled from a uniform distribution ranging from 0.01 to 50%. Linkage disequilibrium (LD) was simulated as described in Yuan *et al.* (2011) starting from an initial population with high LD level, and then decaying to the desired level through the processes of mating and recombination over generations. The case population under the null hypothesis was generated by randomly picking 2000 subjects from the initial population of 4000.

2.4 Type II error simulations

To assess ABACUS power to detect true positives on a benchmark with known genotype–phenotype associations under different conditions, such as different number of SNPs associated with the disease, different MAF and LD patterns across variants, we run a number of different simulations. In particular, we generated 100 datasets of 4000 subjects and 20 000 SNPs and for each dataset, we sampled 1000 SNP-sets of 20 SNPs each, for a total of 10^5 simulated SNP-sets. The LD was simulated as described in the previous paragraph. Each SNP-set had a number of SNPs randomly associated with the disease equal to 1, 2, 3, 4, 6, 8 or 10. The assignment was done having care to produce SNP-sets at low and high MAF, in LD and not, to being able to assess methods performance under different conditions.

After having assigned a phenotype label to each subject (e.g. half controls and half cases), for each associated SNP allele frequencies were distributed in cases according to 10 different models, including single-locus and two-locus interactions, and to different odds ratio sampled in the range 1.4–3. In particular, the model set included single-locus recessive, dominant, additive and multiplicative models and six two-locus models (Fig. 2), where the risk genotypes were all assumed to carry the same risk. All the details on how we assigned allele frequencies accordingly to the different models are given in the Supplementary Material.

2.5 Real data

As a proof of concept, we applied ABACUS on the WTCCC case-control study on T1D and T2D (The Wellcome Trust Case Control Consortium, 2007). The study examined ~2000 T1D cases, 2000 T2D cases and 3000 healthy controls. Each subject was genotyped on the Affymetrix GeneChip 500K Mapping Array Set. We excluded a small number of subjects according to the sample exclusion lists provided by the WTCCC. In addition, we excluded a SNP if (i) it is on the SNP exclusion list provided by the WTCCC and (ii) it has a poor cluster plot as defined by the WTCCC. The resulting dataset consists of 458376 SNPs, measured for 1963 T1D cases, 1924 T2D cases and 2938 controls.

	SNP i		
SNP j	XX	Xx	xx
YY	0	0	0
Yy	0	0	0
yy	0	0	1

	SNP i		
SNP j	XX	Xx	xx
YY	0	0	0
Yy	0	0	0
yy	0	1	1

	SNP i		
SNP j	XX	Xx	xx
YY	0	0	1
Yy	0	1	0
yy	1	0	0

	SNP i		
SNP j	XX	Xx	xx
YY	0	0	0
Yy	0	1	1
yy	0	1	1

	SNP i		
SNP j	XX	Xx	xx
YY	0	0	0
Yy	0	0	1
yy	0	1	1

	SNP i		
SNP j	XX	Xx	xx
YY	0	0	1
Yy	0	0	1
yy	1	1	0

Fig. 2. The six two-locus models implemented in the simulation. The combinations of the two SNPs genotypes that carry higher risk of disease are marked with symbol 1

We mapped SNPs to genes using Affymetrix SNP Array 500K annotation (http://www.affymetrix.com). SNPs annotated as 3′ and 5′ untranslatable region, coding sequences, intron, upstream and downstream were all associated to the corresponding gene. Multiple SNP–gene associations were allowed. We mapped genes to pathways using the Molecular Signatures Database, mSigDB (Subramanian *et al.*, 2005) and exploited the curated SNP-sets derived from the REACTOME and KEGG pathway databases (Kanehisa and Goto, 2000; Matthews *et al.*, 2009). SNPs that did not map to any pathway due to lack of annotation were mapped on 22 pseudo-pathways corresponding to the different chromosomes.

2.6 Implementation of KBAT and SKAT-O

We compared ABACUS with KBAT and SKAT-O. We chose KBAT because it has been reported to outperform other methods under different disease–SNP association models and assumptions (Asimit and Zeggini, 2010; Mukhopadhyay *et al.*, 2010) and SKAT-O because it optimally combines a burden test with a kernel method and is one of the latest proposed approaches by Lee *et al.* (2012).

We run SKAT-O in R using the package SKAT using the optimal adjustment method, optimally compromising between a burden test and the original SKAT algorithm (Wu *et al.*, 2011). The SKAT function in R allows setting two parameters a_1 and a_2, to balance between the weights of rare and more common variants in determining the SNP-set *P*-value (see Wu *et al.*, 2011 for more details on these parameters). We run SKAT-O using both the default parameters setting, $a_1=1$ and $a_2=25$, which weights more the rare variants, and an alternative parameters setting, $a_1 = 25$ and $a_2 = 1$, which weights more the common variants, so to explore both conditions. All the other parameters were set to their default.

We implemented KBAT with the allele match kernel (AM) as explained in Mukhopadhyay *et al.* (2010), as this is the only kernel that does not require knowledge of the risk allele of each SNP.

3 RESULTS

3.1 Ability to control the type I error rate

To investigate whether ABACUS attains to the desired type I error rate at low significance levels, we simulated 10^7 SNP-sets

Table 2. Estimated type I error rates at different significance levels α, for both SNP-set and single SNP association

α	$5{\cdot}10^{-2}$	$5{\cdot}10^{-3}$	$5{\cdot}10^{-4}$	$5{\cdot}10^{-5}$
SNP-set	$1.6{\cdot}10^{-2}$	$1.2{\cdot}10^{-3}$	$1.1{\cdot}10^{-4}$	$0.9{\cdot}10^{-5}$
Single SNP	$8.3{\cdot}10^{-4}$	$5.6{\cdot}10^{-5}$	$5.2{\cdot}10^{-6}$	$4.3{\cdot}10^{-7}$

under the null hypothesis as explained in Section 2.3. Results are shown in Table 2 and suggest that ABACUS is able to control the type I error rate for SNP-set association in a slightly conservative way, probably due to the implementation of the Bonferroni correction on the statistic S_2 within each SNP-set, followed by the graph pruning strategy. However, despite being slightly conservative, ABACUS maintains good power compared with the state-of-the-art methods, as shown later in the text (Section 3.2).

Because ABACUS outputs both the SNP-sets associated to the phenotype and the SNPs within the SNP-sets that contributed to the association, it is of interest to comment also on the false-positive rate for single SNP association. In this regard, it must be noted that ABACUS selects a SNP-set as associated to the phenotype if at least one SNP in the SNP-set is associated with the phenotype. As explained in Section 2.2, ABACUS controls the familywise error rate of these associations at the chosen significance level α within each SNP-set; in other words, the probability of committing at least one false-positive single SNP association within each SNP-set is equal to α. In a dataset with K SNP-sets (let us suppose SNP-sets are statistically independent for the sake of simplicity), the expected percentage of false-positive single SNP associations is thus equal to α multiplied by K, divided by the total number of SNPs being analyzed. In our simulations, this corresponds to α divided by 20. Consistently, the empirical type I error rate observed for single SNP association (Table 2, second row) takes a value close to the type I error rate observed for SNP-set association divided by 20.

3.2 Comparison with KBAT and SKAT-O

One hundred simulated datasets were used to evaluate ABACUS performance under different scenarios; namely, different MAF, different numbers of associated SNPs and presence/absence of LD within the SNP-set being analyzed, in comparison with KBAT and SKAT-O. Each dataset consists of 1000 SNP-sets, thus the false-positive rate was controlled at a significance level α equal to 0.005. Because KBAT and SKAT-O give their output in terms of significance of association of the SNP-set with the disease, to compare the different methods we used the recall, defined as the number of true positive SNP-sets divided by the number of SNP-sets truly associated with the disease.

Figure 3 shows the results averaged across the 100 simulated datasets, for low MAF, i.e. ≤ 0.1 (left panels), and high MAF, i.e. >0.1 (right panels), under single- and multi-locus association models, with associated SNPs in LD (upper panels) or not (lower panels).

As for the other methods, ABACUS performance is robust to different correlation among SNPs and improves with the number of SNPs associated with the disease. A one side Wilcoxon signed

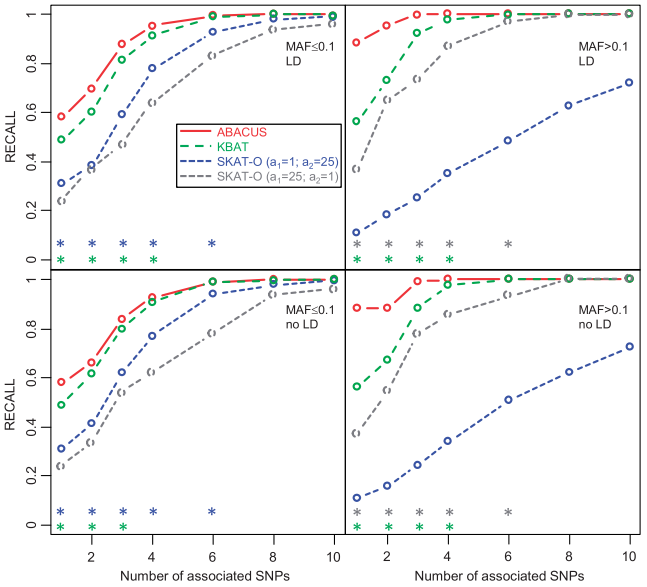

Fig. 3. Recall of different methods in detecting associated SNP-sets. Average recall across 100 simulated datasets is shown at low (left panels) and high (right panels) MAF, for SNPs in LD (upper panels) and not in LD (lower panels), as a function of the number of SNPs associated to the phenotype in the SNP-set. Colored star symbols indicate statistically significant differences between ABACUS and the other tests

ranks test corrected for multiple testing was used to test if ABACUS significantly outperforms other methods ($\alpha = 0.05$). Results are shown in Figure 3, where colored star symbols indicate statistically significant differences between ABACUS and the other tests. Because SKAT-O performance depends on the parameters a_1 and a_2 and, as evident from Figure 3, there is not a unique parameter setting working well for both rare and common variants, we compared ABACUS with the best performing setting of SKAT-O. Results indicate that ABACUS outperforms all other methods in terms of ability to detect true associations when few SNPs in the SNP-set are associated with the phenotype. From Figure 3, it appears also evident that ABACUS and KBAT are robust to different MAFs, with ABACUS improving its performance at high MAFs.

The sensitivity of ABACUS in detecting rare variants even when few SNPs are associated to the phenotype derives from the use of a pairwise statistic S_2 [Equations (1) and (2)]. In fact, differently from other methods, which assign a score and the corresponding P-value to the SNP-set being analyzed, ABACUS assigns a score to each pair of SNPs, which increases with (i) the strength of association, (ii) the number of SNPs associated to the disease and (iii) possible epistatic interactions between SNPs. For example, let us consider a SNP-set of 10 SNPs. In the worst-case scenario, i.e. the most difficult to detect, just 1 SNP, say k, is associated to the phenotype within the SNP-set. In this case, ABACUS computes $45 = 10*9/2$ observed values of S_2, of which 9 involve k. Even in case k corresponds to a rare variant, there is a reasonable chance that at least 1 of the 9 values of $S_2(k, j)$ $(j=1, \ldots, 10, j \neq k)$ passes the threshold θ (step 4 of the algorithm). In case 2 of 10 SNPs are associated with the phenotype, 16 statistics S_2 will involve one of

the SNPs associated with the phenotype and 1 statistic will involve both, thus improving the chance that at least 1 statistic passes the threshold θ. Moreover, if an epistatic interaction between the two associated SNPs exists, the statistic S_2 between them will further increase its value. Obviously, the power of ABACUS keeps increasing with the number of associated SNPs.

3.3 Application to T1D and T2D data

On T1D ABACUS identified 864 SNP associations, corresponding to 153 genes and 267 SNP-sets; on T2D ABACUS identified 75 SNP associations, corresponding to 45 genes and 67 SNP-sets. The MAF distribution of the SNPs identified by ABACUS as associated with T1D and T2D covers the entire range 0.01–0.5 for both datasets (Fig. 4). Redundancy among SNP-sets is obviously observed, as a gene may function in multiple ways and thus may appear multiple times in functional SNP-sets and SNPs can be mapped to multiple genes. In Supplementary Material 'PATH_tables' non-redundant pathways associated with T1D and T2D, i.e. pathways whose selected SNPs were not entirely included in other pathways, are shown, together with the number of selected SNPs and the corresponding genes. The complete lists of associated SNPs and relative information (SNP ID, chromosome, cytoband, MAF, odds ratio and Affymetrix Annotation) are available as Supplementary Materials (RES_T1D.txt and RES_T2D.txt).

As regards T1D, all pathways including genes from the human leukocyte antigen (HLA) region of chromosome 6 (6p22.2 ~ 6p21.32) received high ranking and had 758 SNPs associated among the 864 selected. This was to be expected, as HLA is one of the most recognized regions of interest for T1D (Barrett *et al.*, 2009) and HLA genes, playing a central role in the human immune system, are known vulnerabilities to autoimmune diseases (Altmann and Trowsdale, 1989; Seliger, 2012). Pathways from 1 to 9 and from 16 to 23 in Supplementary Material PATH_tables refer to region 6p22.2 ~ 6p21.32.

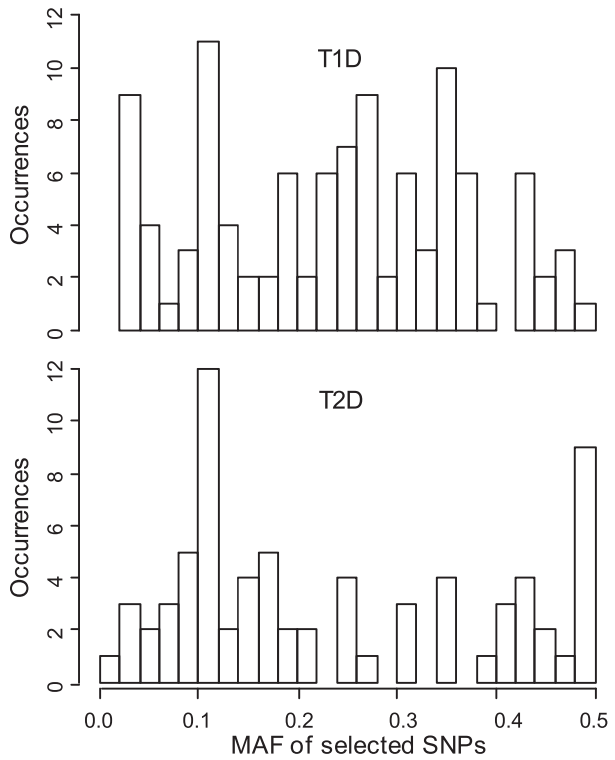

Fig. 4. MAF distribution. MAF distribution of the SNPs identified by ABACUS as associated with T1D (upper panel) and T2D (lower panel)

For both type 1 and type 2 diabetes, ABACUS was able to detect SNPs/genes previously identified in the WTCCC study, plus numerous additional SNPs/genes identified in different T1D and T2D GWAS. In particular, 90% of the SNPs selected by ABACUS were already associated or map in regions of the genome previously associated with diabetes, metabolic traits, LDL cholesterol, body mass index, fasting glucose-related traits and glycated hemoglobin levels. These terms and the related SNP associations were retrieved from the database HuGE Navigator (Wei *et al.*, 2011), searching for human genetic associations with the term 'diabetes' http://hugenavigator.net/HuGENavigator/gWAHitStartPage.do.

The 10% remaining SNPs map on genes and pathways that are biologically sound. We briefly list and comment these pathways in the following.

- 'Focal Adhesion' is crucial for glucose-stimulated insulin secretion (Rondas *et al.*, 2011);
- 'Platelet homeostasis' is central in T1D, as platelet hyperactivity and abnormal Ca(2+) homeostasis in diabetes mellitus are known (Li *et al.*, 2001);
- 'GPCR downstream signaling' and, in particular, the gene adenylyl cyclase 8 are central to glucagon-like peptide 1 signaling (Roger *et al.*, 2011);
- 'Apoptosis' and 'IL receptor SHC signaling' might be of great interest, as apoptosis is involved in T1D beta-cells death and IL3RA is reported to be highly expressed in beta cells in human (source: http://www.t1dbase.org/);
- 'Inositol phosphate metabolism' and, in particular, pyrophosphates have been reported to inhibit Akt signaling and thus insulin sensitivity (Chakraborty *et al.*, 2010);
- 'Glutathione metabolism' has been reported to be altered in adolescents with type 1 diabetes (Darmaun *et al.*, 2008);
- 'Huntigton disease' has been controversely associated with increased risk of T1D. Interestingly, the two genes we have found in association with this pathway might be of interest, as (i) the brain-derived neurotrophic factor has been shown to exert an important role during implantation, placental development and fetal growth and to have low expression when fetal macrosomia is associated with maternal type 1 diabetes (Mayeur *et al.*, 2010) and (ii) the nuclear respiratory factor-1 (NRF-1) enhances the promoter activity of mitochondrial transcription factor A (a key regulator of mitochondrial DNA transcription and replication) at high glucose levels. (Choi *et al.*, 2004);
- 'Tryptophan metabolism' has been previously associated with T2D, as tryptophan levels are low in type 2 diabetic patients, and interestingly, also in gestational diabetic women;

- 'COPI-mediated transport' might be of interest, as Coat protein complex, or COPI complex is a regulator of lipid homeostasis (Beller *et al*., 2008).

It is interesting to note that, for each of the pathways listed above, which refer to regions not previously associated with diabetes, only few SNPs (from 1 to 6) were associated to the phenotype, possibly confirming the sensitivity of ABACUS when few markers in the SNP-set are associated with the phenotype.

Overall, the aforementioned results provide proof of the principle that ABACUS is a powerful tool to identify genotype–phenotype associations within predefined sets of functionally related genes, thus confirming the good sensitivity observed on simulated data. Moreover, the application of ABACUS to biological pathways gives an implicit functional characterization of trait-associated loci.

4 DISCUSSION

In this work, we have presented ABACUS, a method for identifying genotype–phenotype associations within predefined sets of SNPs in GWAS studies. ABACUS is based on the concept of entropy, which allows measuring the variation in the information content of genotype frequencies in cases versus controls. In particular, ABACUS selects the SNPs associated with the phenotype within a SNP-set based on the median value $MS(i)$ of the bivariate statistic $S_2(i, j)$ [(Equations. (1) and (2)], this latter measuring the joint effect of SNPs i and j on the phenotype. It is important to note that the magnitude of $MS(i)$ is SNP-set dependent, its value being enhanced by the presence of different SNPs associated to the phenotype in the SNP-set and by possible epistatic associations between them. However, $MS(i)$ does not depend on the number of SNPs being analyzed in the SNP-set; for example, for the WTCCC T1D and T2D datasets, the Pearson correlation between $MS(i)$ and the number of SNPs in the SNP-set is -0.011 and 0.038, respectively.

The rationale of the iterative graph pruning strategy performed at steps 3–5 of the algorithm is to account for the confounding effect that strongly associated SNPs would have on the statistic $MS(i)$ of the other SNPs in the SNP-set. Steps 3–5 of the algorithm result in a significant improvement in terms of precision in the detection of single SNP associations (number of true positive SNPs divided by the number of selected SNPs), which rises from an average value of 0.9 to 0.98 across the simulated datasets (Wilcoxon paired test $P < 1 \times 10^{-10}$), without a loss of recall.

On simulated data ABACUS has been shown to be robust to different MAF and different correlation among SNPs; moreover, it is more powerful than other methods in terms of ability to detect true associations when few SNPs in the SNP-set are associated with the phenotype. The robustness of ABACUS to different MAF and its sensitivity when few markers are associated with the phenotype is confirmed by the use of real data.

Because ABACUS is able to simultaneously consider common and rare variants and different directions of genotype effect, we consider pathways as the preferred definition of SNP-sets. Focusing on multi-locus associations within a set of functionally related genes might shed light on functional interactions and might represent an advance in the direction of a systems-level understanding of gene regulation. Besides observing a wide overlap with already known associations, analyzing ABACUS results we found a number of biologically sound pathways, which might help generating new hypothesis on diabetes mechanism and etiology.

ABACUS, here described for a two-class problem, can be easily extended to more classes. Moreover, with sufficient numbers of cases and controls or focusing on a genomic region known to be rich of possible interactions, ABACUS can be extended to analyze the joint effect of three or more variables.

In its present form, ABACUS does not output risky combinations of alleles of the output SNPs. However, this can be easily derived from ABACUS output using standard methods and tools such as the logistic regression implemented in PLINK (Purcell *et al*., 2007). One of the possible future directions is to study the combinatorial problem of finding the best partition of susceptibility SNPs in sets of one, two, . . . , k variables, each set with a joint effect on the disease.

ACKNOWLEDGEMENTS

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

Funding: The research was supported by European Union's Seventh Framework Programme (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement n. IMI/115006 (the SUMMIT consortium).

Conflict of Interest: none declared

REFERENCES

Altmann,D.M. and Trowsdale,J. (1989) Major histocompatibility complex structure and function. *Curr. Opin. Imunol.*, **2**, 93–98.

Asimit,J. and Zeggini,E. (2010) Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.*, **44**, 293–308.

Barrett,J.C. *et al*. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, **41**, 703–707.

Beller,M. *et al*. (2008) COPI complex is a regulator of lipid homeostasis. *PLoS Biol.*, **6**, e292.

Chakraborty,A. *et al*. (2010) Inositol pyrophosphates inhibit Akt signaling, thereby regulating insulin sensitivity and weight gain. *Cell*, **143**, 897–910.

Choi,Y.S. *et al*. (2004) Regulation of mitochondrial transcription factor A expression by high glucose. *Ann. N. Y. Acad. Sci.*, **1011**, 69–77.

Darmaun,D. *et al*. (2008) Poorly controlled type 1 diabetes is associated with altered glutathione homeostasis in adolescents: apparent resistance to N-acetylcysteine supplementation. *Pediatr. Diabetes*, **9**, 577–582.

Di Camillo,B. *et al*. (2012) Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. *PLoS One*, **7**, e32200.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

Fan,R.Z. and Knapp,M. (2003) Genome association studies of complex diseases by case-control designs. *Am. J. Hum. Genet.*, **72**, 850–868.

Gibson,G. (2012) Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, **13**, 135–145.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Lee,S. *et al.* (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.

Li,Y. *et al.* (2001) Platelet hyperactivity and abnormal Ca(2+) homeostasis in diabetes mellitus. *Am J Physiol Heart Circ Physiol.*, **280**, H1480–H1489.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Madsen,B.E. and Browing,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Manolio,T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Matthews,L. *et al.* (2009) Reactome knowledgebase of biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.

Mayeur,S. *et al.* (2010) Placental BDNF/TrkB signaling system is modulated by fetal growth disturbances in rat and human. *Placenta*, **31**, 785–791.

Moore,J.H. *et al.* (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26**, 445–455.

Morgenthaler,S. and Thilly,W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, **615**, 28–56.

Morris,A.P. and Zeggini,E. (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–193.

Mukhopadhyay,I. *et al.* (2010) Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet. Epidemiol.*, **34**, 213–221.

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.

Roger,B. *et al.* (2011) Adenylyl cyclase 8 is central to lucagonsn-like peptide 1 signalling and effects of chronically elevated glucose in rat and human pancreatic beta cells. *Diabetologia*, **54**, 390–402.

Rondas,D. *et al.* (2011) Focal adhesion remodeling is crucial for glucose-stimulated insulin secretion and involves activation of focal adhesion kinase and paxillin. *Diabetes*, **60**, 1146–1157.

Sanavia,T. *et al.* (2012) Improving biomarker list stability by integration of biological knowledge in the learning process. *BMC Bioinformatics*, **13** (**Suppl. 4**), S22.

Schaid,D.J. *et al.* (2005) Nonparametric tests of association of multiple genes with human disease. *Am. J. Hum. Genet.*, **76**, 780–793.

Seliger,B. (2012) Novel insights into the molecular mechanisms of HLA class I abnormalities. *Cancer Immunol. Immunother.*, **61**, 249–254.

Shannon,C.E. and Weaver,W. (1963) *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Subramanian,A. *et al.* (2005) SNP-set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

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.

Wang,T. and Elston,R.C. (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.*, **80**, 353–360.

Wei,Y. *et al.* (2011) GWAS integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *Eur. J. Hum. Genet.*, **19**, 1095–1099.

Wessel,J. and Schork,N.J. (2006) Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.*, **79**, 792–806.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yuan,X. *et al.* (2011) Simulating linkage disequilibrium structures in a human population for SNP association studies. *Biochem. Genet.*, **49**, 395–409.

Zeggini,E. and Asimit,J.L. (2011) An evaluation of power to detect low-frequency variant associations using allele-matching tests that account for uncertainty. *Pac. Symp. Biocomput.*, **2011**, 100–105.