Title Page

Increasing the Power of Association Studies by Imputation-based Sparse Tag SNP Selection

Ofir Davidovich[*,a], Gad Kimmel[*,b], Eran Halperin[a,b], Ron Shamir[a]

offird@gmail.com   gkimmel@gmail.com

heran@post.tau.ac.il   rshamir@tau.ac.il

[a]Blavatnik School of Computer Science, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

[b]International Computer Science Institute, Berkeley, CA 94704, USA.

Corresponding author: Shamir, R. (rshamir@tau.ac.il)

[*]These authors contributed equally to this paper.

Running title: The power of imputation-based tag SNP selection

Abstract

In low- and medium-budget association studies, a limited number of tag SNPs are selected out of a large set of available SNPS previously typed in an initial cohort. These tag SNPs are then typed in a larger set of control and affected individuals. Current association studies pick the set of tag SNPs based on the correlation criterion. Here we show that association studies that use tag SNPs selected according to their imputation accuracy are more powerful than those relying on tag SNPs selected by the correlation criterion. The advantage is particularly striking when the set of tag SNPs is sparse; thus, picking tag SNPs to maximize the imputation accuracy will increase the effectiveness of future association studies without additional cost.

Introduction

Tag SNP selection and association studies

Over the past few years, numerous disease association studies, both genome-wide and of candidate genes, have been undertaken (e.g. Refs [1-7]) and more are expected. The goal of such studies is to identify the genetic causes of complex diseases, by associating the disease phenotype with genetic variants such as single nucleotide polymorphisms (SNPs) or copy number variants (CNVs). The statistical significance and power of a study are directly affected by the number of individuals tested and by the number of SNPs typed: the more SNPs typed, the more information obtained, and the greater the chance that a causal SNP will be among the genotyped SNPs or highly correlated with one of them. However, typing too many SNPs is a waste of resources because of the strong dependency between SNPs that lie close to each other owing to the phenomenon of linkage disequilibrium (LD) [8]. Although the cost of genotyping has dropped dramatically in recent years, the cost of customized genotyping platforms remains high, and budgetary constraints prevent many laboratories from genotyping a dense set of SNPs in the target regions. Ideally, one would like to select and type a given number of SNPs in a way that maximizes the power of the association study (at a fixed significance level). This is usually done by choosing a subset of the SNPs, called *tag SNPs* (or simply *tags*), whose number is commonly determined by the resources available.

Identifying an optimal set of tag SNPs of minimum size with sufficiently high power is germane to several aspects of genome association studies: candidate gene studies (for example see Refs. e.g., [1-5]) seek association around a focused number of genes

rather than performing a genome-wide or unbiased search; low-budget association studies search for a limited number of tag SNPs in suspected regions or genes; and multi-stage studies involve custom chips in the late stages, relying on customized genotyping technologies that are still widely employed (sold by such companies as Taqman of ABI, Sequenom, and the Golden Gate platform of Illumina).

Most current studies use the correlation coefficient as a criterion for the choice of tag SNPs. This approach has certain theoretical justification, but the power of subsequent association studies that use it has not been systematically investigated. Here we show that using imputation accuracy in the tag selection stage gives more power to association studies than the widely used correlation (r^2) measure, especially when sparse tags are sought.

Tag SNP selection criteria

There are many methods for selecting a desirable set of tag SNPs, e.g. [9-14]. Generally, these methods use a training dataset to study the properties of the correlation structure in the genome, and then use that dataset to choose a set of tag SNPs based on some criterion. Criteria vary between methods. In 2005 de Bakker et al. [14] described a method called *Tagger* for tag SNP selection based on the r^2 criterion. r^2 is the square of the correlation coefficient between two SNPs, when the two SNPs are viewed as binary vectors. Formally, for a pair of SNPs with major allele 0 and minor allele 1, $r^2 = \dfrac{(p_{00} - p_0 q_0)^2}{p_0(1-p_0)q_0(1-q_0)}$, where p_{00} is the frequency of the haplotype 00, and p_0, q_0 are the frequencies of the major alleles in the two SNPs. Tagger uses a heuristic greedy algorithm (originally suggested by [11]) to select tags that capture all SNPs with an r^2 that exceeds a prescribed threshold (e.g $r^2 > 0.8$). In

addition, to improve efficiency, it can perform an aggressive search aimed at replacing each tag with a specific multi-marker predictor (on the basis of the remaining tags). Tagger is implemented in the Haploview software [15] developed as part of the HapMap project (www.hapmap.org).

Although Tagger is widely used and has been shown to considerably increase the power of the association studies, its superiority over alternative methods with regard to the power of association tests has not been fully investigated yet. Tagger's rationale for picking the set of tag SNPs according to the correlation coefficient rests on theoretical foundations; specifically, it can be shown that if SNP A is a tag of SNP B, and B is the single causal SNP, then the power to detect association in B by typing A in n individuals is equivalent to typing B directly for $r^2 n$ individuals [8]. Note that Tagger is not guaranteed to find an optimal set of tags.

A natural alternative criterion for tag SNP selection is the prediction accuracy, or *imputation accuracy*, which directly evaluates the average SNP imputation quality, i.e., the average error rate in the imputation of the hidden SNP. This criterion has been originally suggested by Stram [16]. A variant of the imputation accuracy criterion has also been suggested by He and Zelikovsky [17], where multiple linear regression is used to infer the hidden SNPs. Halperin et al. [13] suggested an algorithm called STAMPA, which optimizes the selection of tag SNPs under the imputation accuracy criterion. Under the approximate assumption of monotone LD decay (by which the LD between a pair of SNPs is a monotone function of their physical distance), STAMPA was shown to find an optimal set of tag SNPs using a dynamic programming algorithm. The basic idea behind the STAMPA algorithm is that under LD decay, it is enough to require that each hidden SNP will be imputed by its physically closest tag SNPs from each side. For each hidden SNP and for each set of

potential tag SNPs, we can use the training dataset to estimate the prediction error rate for that SNP. We can then use dynamic programming to search for a set of tag SNPs that minimizes the average prediction error rate.

Although intuitively imputation accuracy is a reasonable optimization criterion, the power of an association study that uses the tags is a more meaningful optimization criterion, because it directly measures the chance of success of the study. It is therefore important to evaluate the relationship between the criterion for tag SNP selection and the power of the association studies that use the tags. Because the power gauges the chance of success of an association study, selection of tag SNPs that yield higher power will increase the effectiveness of future association studies, at no additional expense.

Methods

To empirically test the relation between the tag SNP selection criterion and the power, we directly compared the power of Tagger and STAMPA; to do so, we conducted extensive simulations similar to de Bakker et al. [14]. The outline of the procedure is as follows (Figure 1): Starting from real haplotypes, we generated many case-control panels, each with a known causal SNP. In each panel, after selecting the tag SNPs, we computed the power of the sub-panel containing the tags only, to detect association with the hidden causal SNP. This procedure was repeated for different tag selection methods and for different tagging scenarios.

We evaluated the power obtained by the imputation accuracy criterion for tag SNP selection used by STAMPA (as implemented in the software package GEVALT [25]), to the power obtained by the correlation coefficient criterion used by Tagger (as

implemented in Haploview [15]). As a baseline, we compared the power of these two criteria to a naive algorithm, *Rand,* which randomly selects tag SNPs. Tagger was tested in both the standard mode and in its "aggressive mode", which uses haplotypes (we call the latter here *Tagger-A*).

We now describe the procedure in detail.

1. **Generating panels:** We extracted from the HapMap resource (release 21) phased haplotypes of chromosome 1 of the 60 individuals (parents only) of west European ancestry selected by the Centre d'Etude du Polymorphisme Humain (CEPH) [18]. SNPs with minor allele frequency < 5% were removed, and the dataset was partitioned into 100 disjoint regions of 1,000 contiguous SNPs per region (spanning the first 172Mb of chromosome 1). The average distance between adjacent SNPs was 1,364bp. We selected a causal SNP at random in each region and constructed a panel of 1000 cases and 1000 controls according to a multiplicative model with genotype relative risk of 1.5, using HAPGEN software [19]. We repeated this process 50 times, generating 50 panels per region. Similar disease models are widely used in the literature, and in particular. Tagger was originally evaluated based on a similar disease model. The choice of the relative risk value is somewhat arbitrary, however, one would expect that qualitatively similar results will be obtained when applying the same analysis with a different value.

2. **Selecting tags:** In a separate procedure, we selected tag SNPs in each region using each method. The same number of tag SNPs was selected by each algorithm. The 120 HapMap haplotypes were taken as the training set for STAMPA and Tagger.

3. **Testing association:** For each panel, for the tag SNPs selected by each

To compute association, $\chi 2$ allelic tests with one degree of freedom were computed and the p-values of association were computed using a permutation test (done with RAT software [26]). For STAMPA, the set of tag SNPs was used to predict the values of the other SNPs (as described in [13]), and the entire set of SNPs was tested for association. For Tagger-A, the set of tests was generated as described in [14]. Since Tagger-A uses haplotypes in the case-control panel to generate the tests, we also tested a version of STAMPA that uses haplotypes instead of genotypes in the imputation, for the purposes of comparison. For Tagger and Rand, the $\chi 2$ tests were applied only to the tag SNPs.

In total, 5,000 case-control panels were generated, and 170,000 association tests were performed.

4. **Computing power:** We evaluated the performance of the methods by their *relative power*, computed as follows. For a fixed significance level p_0, we counted the number of panels in which the algorithm had a significant ($\leq p_0$) p-value (calculated as detailed in 4 above). We then divided this number by the number of panels that were detected as significant when all SNPs (including the hidden SNPs) were available. Hence, this measure evaluates how well an algorithm performs in comparison to an algorithm that has full genotype (or haplotype) information. To test whether the difference in relative

Results

Imputation accuracy confers higher power to association studies

Several tagging scenarios were explored. The first scenario used Tagger's default parameters ($r^2 \geq 0.8$) to capture the entire set of SNPs. We then used STAMPA to pick a set of tag SNPs of the same size as the one that was given by Tagger. On average, 26.4% (standard deviation (SD) 6.3%) of the SNPs were selected as tags (an average of one tag SNP per 5.4Kb). Results are shown in **Supplementary Figure 1A**. Both STAMPA and Tagger obtained a high relative power (96-98% under significance levels ranging from $P = 0.05$ to $P = 0.001$) with non-significant differences between the two methods. Random tag selection (*Rand*) had significantly lower power than both, but, interestingly, it attained a high relative power of 90.5-95.5%.

We next applied Tagger's aggressive method (using 2- and 3-marker haplotypes, and default parameters $r^2 \geq 0.8$, LOD threshold 3.0) to capture the entire set of SNPs (see **Supplementary Figure 1B**). The average fraction of SNPs selected as tags was 21.2% (SD 5.3%; i.e., an average of one tag SNP per 6.7kb). The high relative power of both methods remained (95.3-98%). However, STAMPA had a significant advantage over Tagger for p-value cutoffs 0.001 (97% vs. 95.3%, p-value of 8.9×10^{-5}) and 0.005 (97.4% vs. 96.8%, p-value of 0.04).

The density of tag SNPs needed by Tagger to capture all the SNPs in a region was quite high, ~25%. We next explored scenarios where the tag density was lower. We used Tagger's 'best N' method, which, for a pre-specified N, seeks the N tags that capture the maximum number of SNPs with the prescribed r^2 threshold. We set N to be 34, 45, 68, and 136, which corresponds to an average of one tag SNP per 40, 30, 20, and 10 Kb, respectively. We tested both Tagger and Tagger-A (with default parameters $r^2 \geq 0.8$, LOD threshold 3.0). Results are shown in **Figure 2**. All algorithms showed a linear decrease in power as the tag density decreased, but the rate of decrease in STAMPA's performance was much lower than Tagger's. In all tests STAMPA attained a higher power than Tagger but with varied differences. The advantage in the relative power was 11% at a tag distance of 40Kb. STAMPA's advantage is smaller but statistically significant also at tag distances of 20 and 30 Kb under all significance levels (P-values < 0.007), but not at 10 Kb. Interestingly, Tagger-A performed essentially the same as Tagger. When haplotypes instead of genotypes were used in the imputation, STAMPA's advantage was greater, and the differences were statistically significant at all tag distances.

An alternative to Tagger's 'best N' method is to lower the r^2 threshold until the desired number of tags is selected. We applied this method to select tags at tag distances 10 and 40kb and performed the same simulations as above. The tags selected by the alternative method gave uniformly less power (results not shown).

Note that STAMPA maintains acceptable power even when the selected tags are sparse. For example, at significance $p \leq 0.05$, STAMPA's relative power is 86% when the tag distance is 40kb.

The advantage in power is due to imputation accuracy

Our extensive simulations with real haplotypes showed that STAMPA, which optimizes the imputation accuracy, attained significantly and consistently greater power than Tagger, which optimizes r^2, and that this advantage increased at lower densities. This advantage in the power is explained primarily by the imputation accuracy criterion STAMPA uses to select the tags. An alternative explanation ascribes the advantage to the performance of the two algorithms: STAMPA's algorithm selects an optimal set of tags, according to its criterion, whereas Tagger, being a heuristic, may select only a suboptimal set according to its criterion. To test this possibility, we calculated the r^2 coverage of tags selected by each method: For each threshold θ, we calculated the fraction of non-tag SNPs that are captured by some tag with $r^2 \geq \theta$. **Figures 3A,B** show that Tagger had a higher r^2 coverage. Hence, STAMPA had higher power although its tags had lower r^2 coverage, ruling out the latter explanation.

STAMPA was designed to select tags that optimally predict the non-tags. Unlike Tagger, the use of the predicted non-tags in association tests is an essential part of STAMPA. Is STAMPA's advantage due to the different tags chosen, or is it merely because STAMPA predicts the values of the non-tag SNPs and Tagger does not? To answer this question, we selected tags using Tagger and used them to predict the hidden SNPs based on STAMPA's imputation method. We then computed the association for each tag and for each predicted non-tag. The results for the 'best *N*' scenario can be seen in **Figure 2**. The imputation algorithm does improve Tagger's relative power (all P-values < 0.05 using McNemar's test), but only by a modest factor (average increase of 1.1%), and still well below that of STAMPA.

It is possible that STAMPA's advantage is not because of the tag selection method,

but because the imputation method and the criterion match, while they do not match when using Tagger and then STAMPA's imputation method. To test this, one can select tags by each method and impute the non-tags by an unrelated imputation method, such as the widely used method IMPUTE [19]. We applied this procedure with STAMPA, Tagger, and Rand and then computed the correlation coefficient between the imputed SNPs and their real values. As shown in **Supplementary Figures 2 A,B**, STAMPA's selected tag SNPs had a clear advantage over Tagger's when the tag SNPs were at a distance of 30kb or more (Kolmogorov-Smirnov test p-values. For example, at an average tag distance of 40kb, 93% of the imputed SNPs had correlation larger than 0.9 with some STAMPA's tags, and 82% with Tagger's. For distances < 30kb, STAMPA still had significant advantage, but the difference was small.

Imputation accuracy predicts the non-tags better

We also compared the ability of each method to predict the hidden SNPs. For each set of tags, we predicted the hidden SNPs in the original 120 HapMap haplotypes and calculated the r^2 between each real non-tag and its imputation. We note that Tagger is not designed to optimize the imputation accuracy, and therefore our test is meant as a validation that STAMPA does achieve higher imputation accuracy in practice. **Figure 2** illustrates the large advantage of STAMPA over Tagger. The large difference was manifested by the error rates when imputation was applied on the simulated panels. The error rate is defined as the fraction of wrongly predicted alleles. At tag distances of 10, 20, 30, and 40kb, the average error rates of STAMPA were 3.9, 5.2, 6.2, and 7.1%, respectively, while Tagger's error rates were almost twofold higher: 8.9, 9.4, 10.5, and 11.5%, respectively.

Using haplotypes increases the power

As shown above, using haplotypes instead of genotypes in the imputation increased STAMPA's advantage in power over Tagger (see **Figure 1**). This suggests that in real scenarios, if experimental information on phasing is not available (e.g. from trios), one might gain power by performing the association test on haplotypes that have been phased computationally. We could not test this scenario in our simulations because of the added computational burden of phasing many thousands of panels. When we performed phasing using fastPHASE [21] on a limited number of examples and measured the error rate of the phasing algorithm, the switch error rate [21] was small at tag distances of ≤10kb (2-4%) (data not shown). Therefore, the results described above with computationally derived haplotypes are expected to be only marginally less accurate than were we to use the true haplotypes. Hence, even at high tag densities, STAMPA probably has an advantage over Tagger if phasing of the tags is performed before predicting the hidden SNPs.

Discussion

Finding an optimal set of tags remains extremely important for two reasons: First, certain platforms (such as Illumina's Golden-Gate, ABI's TaqMan, Sequenom, and Affymetrix's Custom SNP Kits) are designed for customized genotyping, and require ad-hoc tag selection for each new study. The availability of such a variety of commercial products attests to their ongoing use. Many focused association studies cannot use standard arrays and have to select tags for specific genomic areas, often at modest density owing to budget constraints (this need is also demonstrated by the more than 300 studies using Tagger to date). Second, each new generation of standard

high-density chips involves the selection of a new set of tag SNPs, so future chips produced by these technologies would benefit considerably from more powerful selection methods. Although average tag densities in such chips will increase, there are sparse genomic regions showing low LD [27], on which imputation-based tag selection will have an advantage.

Recently, several studies [19,22-24] used algorithms for imputing untyped SNPs based on typed tag SNPs. The idea was to use all the imputed SNPs when performing the association test. This approach has higher power than using the tags without the predicted SNPs, providing further for the use of tag selection algorithms based on imputation accuracy. Since an imputation procedure is performed on the tags, and only then the association is tested, it makes sense to choose the tags that impute the rest of the SNPs with minimal error rate. As we demonstrated, STAMPA imputes consistently better than Tagger (**Figures 3C,D**). Hence, our results indicate that if imputation accuracy is used instead of r^2 to select the tags, methods that use the imputed SNPs in computing association (e.g., Halperin et al. [13], and Stram [16]) will achieve higher power. Note that the fact that Tagger has been shown to outperform other correlation-based tag SNP selection methods strengthens our confidence in the suggested conclusion; however, since there is no guarantee for the optimality of Tagger, there is always the possibility that an improved algorithm for the optimization of the correlation criterion will result in different conclusions.

Because the same haplotypes used to select tag SNPs were also used to create the case/control data, there is a possible risk of overfitting in our results. Owing to the scarcity of large-scale published haplotypes, we could not completely separate the data for the training and the test phases. As a partial remedy to this problem, we created the test panels with the HAPGEN software [19], which simulates case and

control individuals conditional upon a set of known haplotype data using an estimate of the fine-scale recombination rate across a region. This approach uses the Li-Stephens model [20], and is preferable to a direct resampling approach [14] that produces new haplotypes that are copies of the original HapMap haplotypes. Even if there is a modest amount of overfitting in the reported power levels, because the same simulated datasets were used by STAMPA and Tagger, the qualitative differences between the performances of the two methods should remain the same. Every association study is preceded by selection of the tag SNPs to be typed, whether explicitly by the researchers or by chip manufacturers. The success of an association study critically depends on its power. Here, we demonstrated that using the imputation accuracy in the tag selection stage is preferable to the widely used r^2 measure. We assumed that association is tested on each SNP separately, which is the common practice. However, different measures for tag SNP selection might give higher power in other testing approaches. Although STAMPA might not give optimal power, it outperforms the standard method currently in use. Therefore, the connection between tag SNP selection and power deserves further exploration.

Acknowledgments

References

[1] Gal, I. et al. (2006) A specific RAD51 haplotype increases breast cancer risk in Jewish non-Ashkenazi high-risk women. Eur. J. Cancer 42(8), 1129–1134

[2] Koren, M. et al. (2006) ATM haplotypes and breast cancer risk in Jewish high risk women. Br.J. Cancer 94, 1537–1543

[3] Pharoah, P. D. P. et al. (2007) Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genetics 3(3):e42

[4] Choudhury, K. et al. (2007) A Genetic Association Study of Chromosome 11q22−24 in Two Different Samples Implicates the FXYD6 Gene, Encoding Phosphohippolin, in Susceptibility to Schizophrenia. Am J Hum Genet 80, 664–672

[5] Di Bernardo MC et al (2008) A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia, Nat Genet,. 40(10):1204-10

[6] 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, 661-683

[7] Gudbjartsson DF et al., (2008) ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma, Nat Genet., 40(8):1029.

[8] Pritchard, J. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69, 1–14

[9] Zhang, K. et al. (2002) A dynamic programming algorithm for haplotype block partitioning. Proc. Natl. Acad. Sci. USA 99(11), 7335–9

[10] Bafna, V. et al. (2003) Haplotypes and informative SNP selection algorithms: Don't block out information. In Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 03), 19–27. The Association for Computing Machinery

[11] Carlson, C. S. et al. (2004) Selecting a maximally informative set of single nucleotide polymorphisms for association analysis using linkage disequilibrium. Am. J. Hum. Genet. 74, 106–120

[12] Pe'er, I. et al. (2004) A branch and bound algorithm for the chromosome tagging problem. In Proceedings of the Second RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotypes, 89–98

[13] Halperin, E. et al. (2005) Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics 21, Suppl. 1, i195–i203

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

[15] Barrett, J. C. et al. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–265

[16] Stram, D. O. (2004), Tag SNP selection for association studies, Genetic Epidemiology 27 (4) : 365-374

[17] He, J. and Zelikovsky, A. (2006), MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression, Bioinformatics 22(20):2558-2561.

[18] The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437, 1299–1320

[19] Marchini, J. et al. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39, 906–913

[20] Li, N. and Stephens, M. (2003) Modelling linkage disequilibrium and identifying recombinations hotspots using SNP data. Genetics 165, 2213–2233

[21] Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large−scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78, 629–644

[22] Zaitlen, N. et al. (2007) Leveraging the HapMap Correlation Structure in Association Studies. Am J Hum Genet 80, 683–691

[23] Servin, B. and Stephens, M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genetics 3(7): e114. doi:10.1371/journal.pgen.0030114

[24] Pe'er, I. et al. (2006) Evaluating and improving power in whole genome association studies using fixed marker sets. Nat Genet 38(6), 663–667

[25] Davidovich, O. et al. (2007) Gevalt: An integrated software tool for genotype analysis. BMC Bioinformatics 8:36

[26] Kimmel, G. and Shamir, R. (2006) A fast method for computing high significance disease association in large population-based studies. Am. J. Hum. Genet. 79, 481–492

[27] Shifman, S., et al. Linkage disequilibrium patterns of the human genome across populations Human Molecular Genetics 12, 771-776 (2003)

Figures
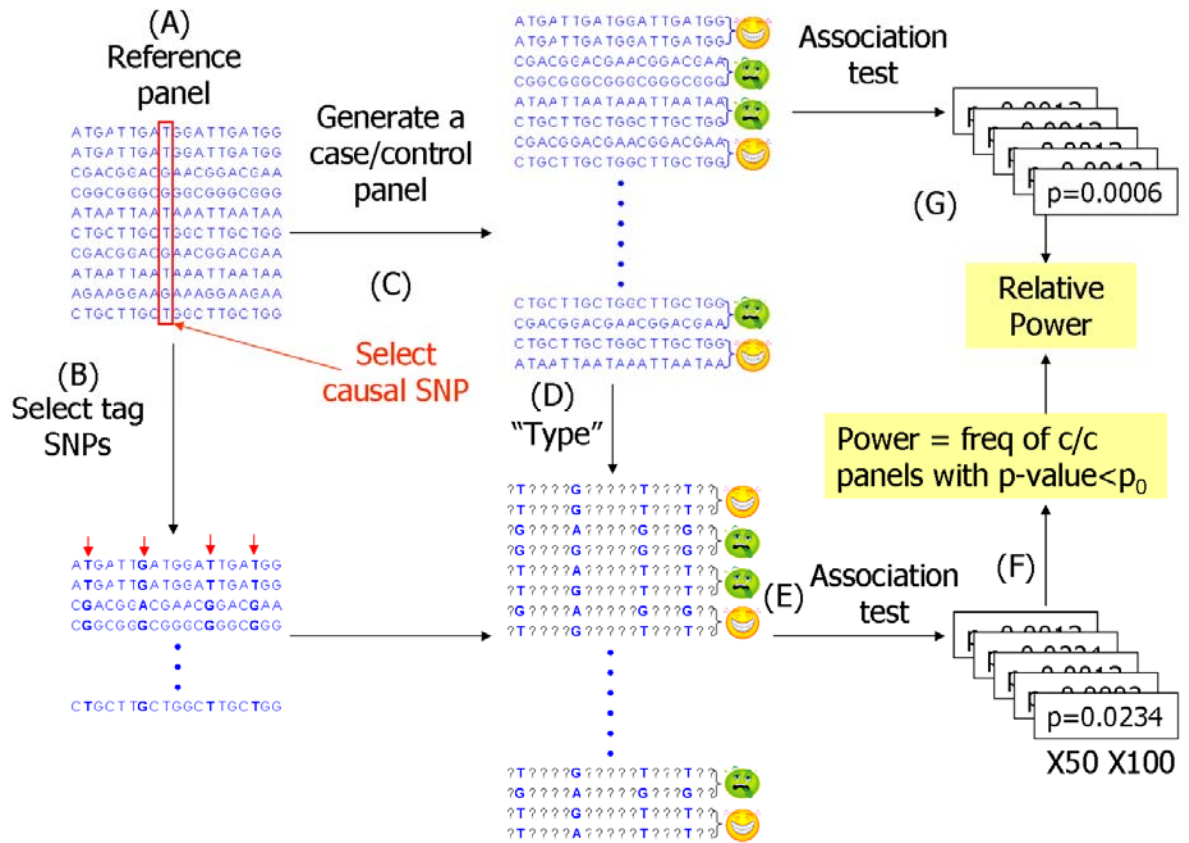
FIGURE 1 Overview of the testing procedure. The procedure for one test panel in one region and for one tag selection algorithm is shown. (A) We start with a reference panel containing the 120 HapMap haplotypes. (B) Tag SNPs are chosen. (C) A causal SNP is randomly selected and a panel of 1000 cases and 1000 controls is simulated. (D) The non-tag SNPs are hidden. (E) The panel of case/control genotypes is tested for association. (F) Steps C-E are repeated 50 times. The whole procedure is repeated in 100 genomic regions, resulting in a list of 5000 p-values from which the power of the tag SNP selection algorithm is calculated. (G) Each case/control panel is also tested for association without hiding the non-tag SNPs. This allows calculation of the relative power of the algorithm.
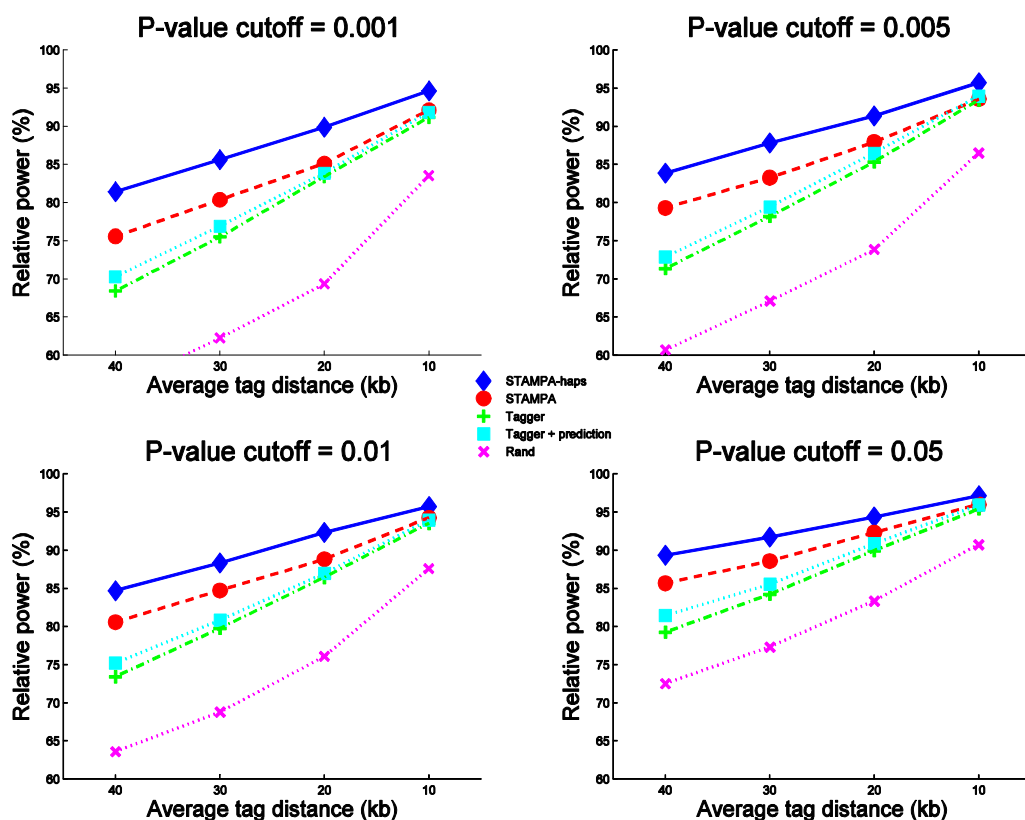
FIGURE 2 Relative power at different tag SNPs densities. The relative power is shown for STAMPA, Tagger, Rand, STAMPA using haplotypes, and Tagger using prediction. The number of tag SNPs per tested region was limited to 34, 45, 68, and 136, which corresponds to an average distance between tags of 40, 30, 20, and 10kb, respectively. The power is shown in four different p-value cutoffs. Tagger+prediction selects tags using Tagger and then predicts the non-tags using STAMPA's method. Tagger-A gave virtually identical results to Tagger (not shown).
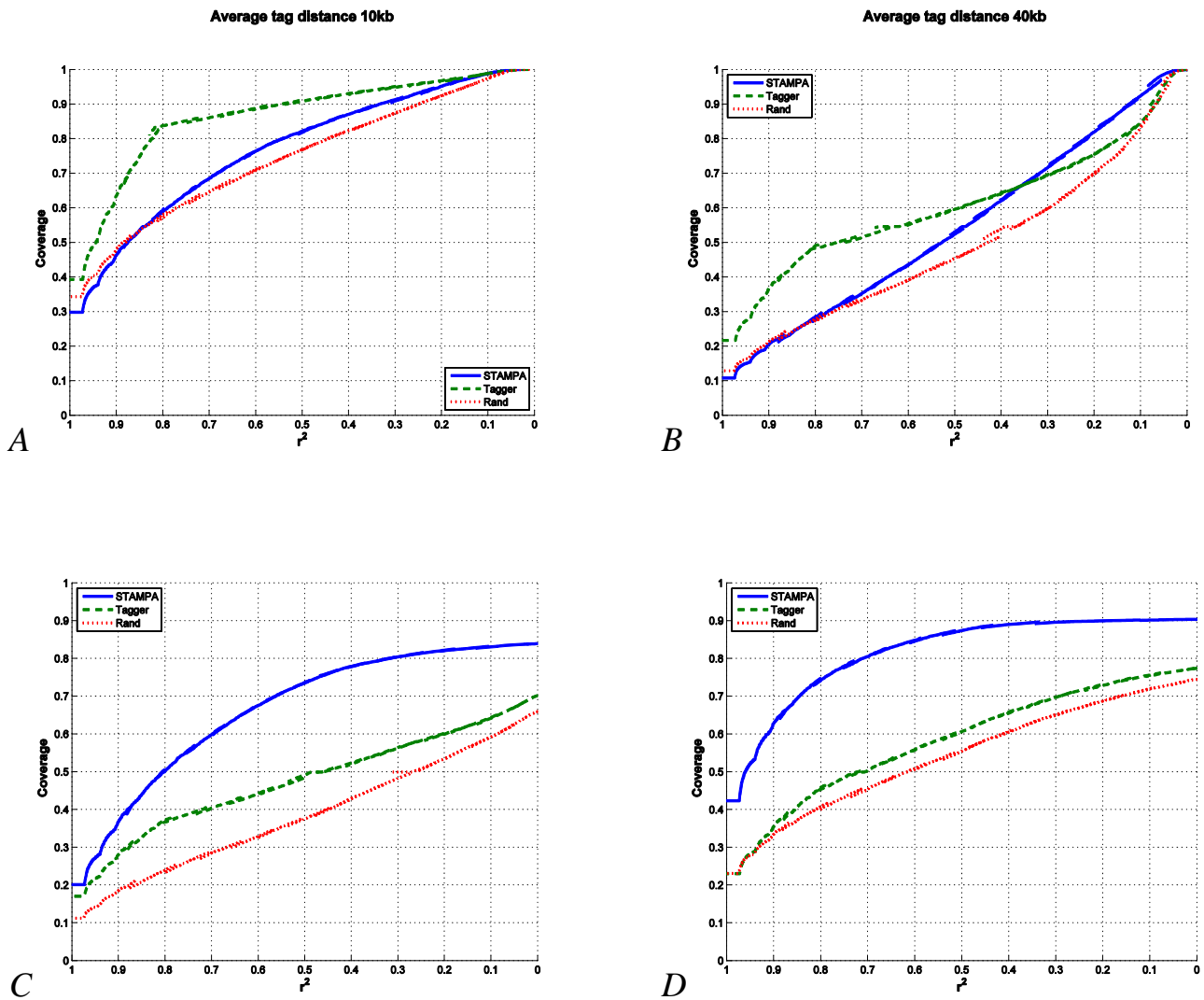
FIGURE 3 Average SNP coverage as a function of the r^2 threshold. A,B: For every value x of r^2 the plots show the fraction of non-tag SNPs that have $r^2 \geq$ x with some tag SNP. C,D: For every value x of r^2 the plots show the fraction of non-tag SNPs that have $r^2 \geq$ x with their predicted values as computed on the HapMap haplotypes. The curves do not reach 100% since some of the SNPs are predicted to be monoallelic, which makes the r^2 measure undefined. Results are shown for the datasets with 10kb (A,C) and 40kb (B,D) tag distance. Tagger-A gives almost identical results to Tagger (not shown).
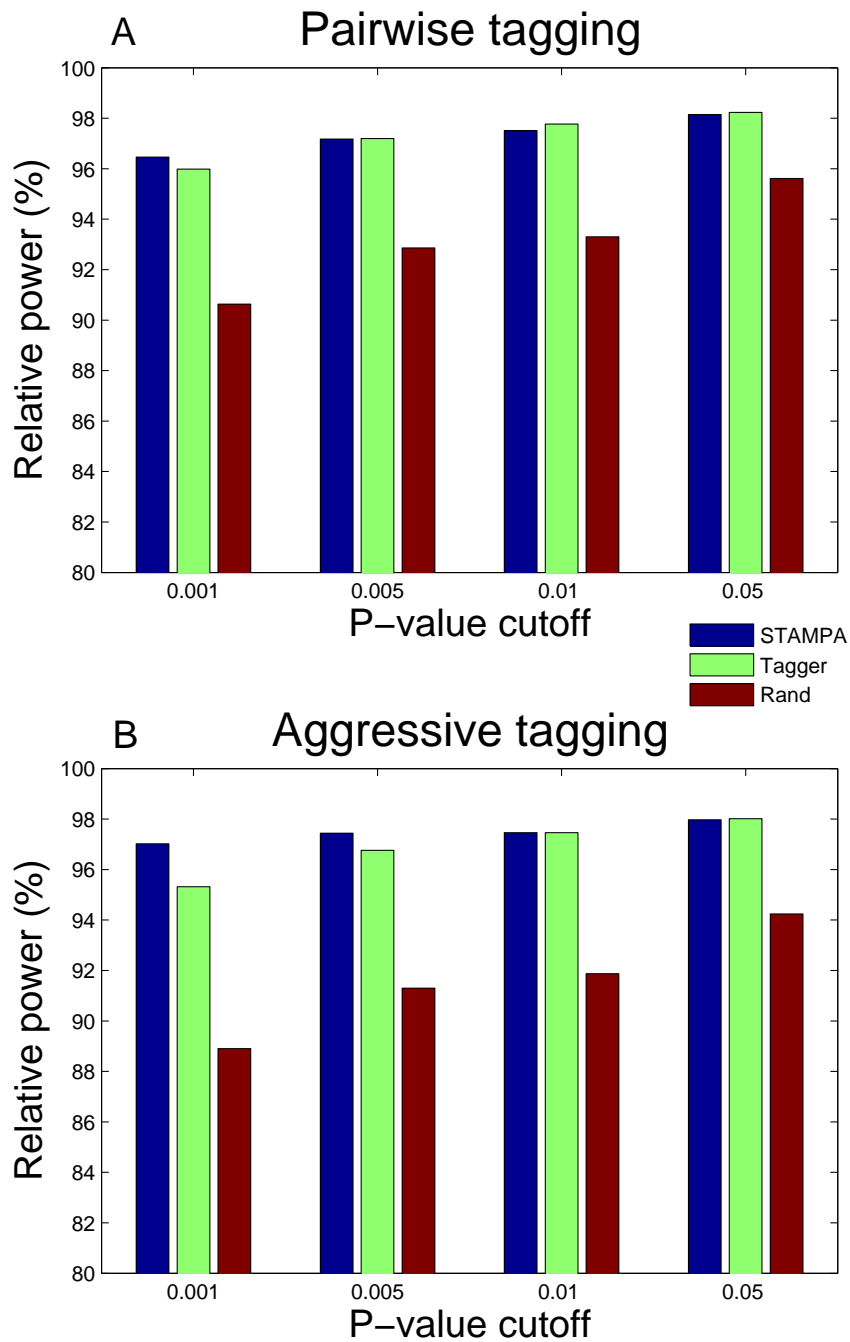
SUPPLEMENTARY FIGURE 1 Relative power of the three methods: STAMPA, Tagger and Rand, for different significance levels. A: using pairwise tagging. B: using aggressive tagging on haplotypes.
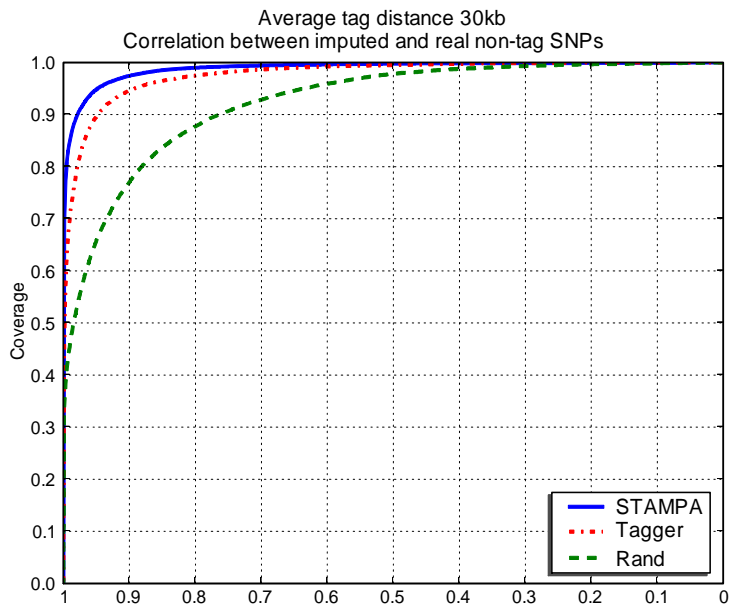
A
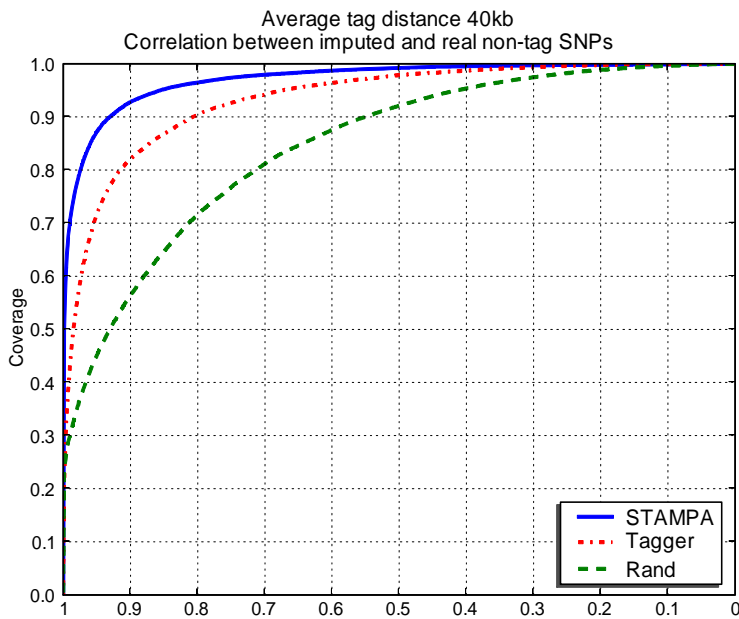
SUPPLEMENTARY FIGURE 2 Correlation coefficient between the imputed SNPs

and their real values. The tag SNPs are selected by each method - STAMPA, Tagger,

and Rand - and the imputation of the non-tags is performed by the IMPUTE

algorithm. For every value x of r^2 the plots show the fraction of non-tag SNPs that

have $r^2 \geq x$ with their predicted values as computed on the HapMap haplotypes. A:

Average tag distance of 30kb, B: Average tag distance of 40kb.