Trick or treat: The effect of placebo on the power of pharmacogenetic association studies

Clara Singer,[1] Iris Grossman,[1,2] Nili Avidan,[1] Jacques S. Beckmann[1,3] and Itsik Pe'er[1]*

[1] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel
[2] Division of Neuroimmunology and MS Center, Rappaport Faculty of Medicine and Research Institute, Technion, Haifa, Israel
[3] Department of Medical Genetics, CHUV-Université de Lausanne, Lausanne, Switzerland
* *Correspondence to*: Tel: +41 21 314 3378; Fax: +41 21 314 3392; E-mail: Jacques.Beckmann@hosprd.ch

Date received (in revised form): 9th February 2005

Abstract

The genetic mapping of drug-response traits is often characterised by a poor signal-to-noise ratio that is placebo related and which distinguishes pharmacogenetic association studies from classical case-control studies for disease susceptibility. The goal of this study was to evaluate the statistical power of candidate gene association studies under different pharmacogenetic scenarios, with special emphasis on the placebo effect. Genotype/phenotype data were simulated, mimicking samples from clinical trials, and response to the drug was modelled as a binary trait. Association was evaluated by a logistic regression model. Statistical power was estimated as a function of the number of single nucleotide polymorphisms (SNPs) genotyped, the frequency of the placebo 'response', the genotype relative risk (GRR) of the response polymorphism, the strategy for selecting SNPs for genotyping, the number of individuals in the trial and the ratio of placebo-treated to drug-treated patients. We show that: (i) the placebo 'response' strongly affects the statistical power of association studies — even a highly penetrant drug-response allele requires at least a 500-patient trial in order to reach 80 per cent power, several-fold more than the value estimated by standard tools that are not calibrated to pharmacogenetics; (ii) the power of a pharmacogenetic association study depends primarily on the penetrance of the response genotype and, when this penetrance is fixed, power decreases for larger placebo effects; (iii) power is dramatically increased when adding markers; (iv) an optimal study design includes a similar number of placebo- and drug-treated patients; and (v) in this setting, straightforward haplotype analysis does not seem to have an advantage over single marker analysis.

Keywords: *trial design, single nucleotide polymorphism, haplotype, power, simulation*

Introduction

Pharmacogenetics (PGx) — the study of how genetic differences influence the variability in patients' response to drugs[1] — investigates genes ideally covering all of the drug's interactions in the course of its passage through the body.[2] The objective of PGx research is to identify the genetic profile contributing to an individual's response pattern to a specific drug. Little is known about the genetic basis of differential drug response. There are examples where a single gene may exert a dominant effect on treatment efficacy, as in the case of cytochrome P4502D6 (CYP2D6), where deficient patients need to be identified before treatment initiation by codeine and its derivatives due to efficacy loss.[3] More commonly, the phenotype of drug response is classified as multifactorial, as it generally results from the interaction of a number of different genetic, as well as environmental, factors. An example of this is the efficacy of clozapine therapy in the treatment of schizophrenia.[4]

Traditionally, genetic mapping can be approached either by linkage (family-based) methods or by association study (population-based) designs. The latter are particularly likely to play a prominent role in pharmacogenetics, as it may be difficult to collect informative families with multiple patients treated with the same drugs. The simplest and most widely applied strategy of association studies is the case-control design; however, several key aspects distinguish PGx association studies from standard disease-oriented case-control studies. First, PGx association studies are usually based on either prospective or ongoing clinical trials, where, classically, patients are randomly assigned to one of two groups: a treatment group, receiving the tested drug; and a control group, receiving placebo (randomised, controlled study). As a result, the number of responders ('cases') can only be determined once the study has been completed and not *a priori*, complicating the recruitment of the required cohort. Secondly, PGx association studies in general, and those of medications for

psychiatric and immunological diseases in particular, are characterised by a poor signal-to-noise ratio: approximately one third of the patients enrolled in efficacy trials may respond to placebo treatment. The placebo 'response' in randomised clinical trials includes such statistical artefacts as regression to the mean,[5] drift in measurement of the response over time and bias of expectations by both patients and evaluators, as well as real effects such as spontaneous recovery, a tendency to seek treatment outside the study and the response to additional attention and concern arising from participation in clinical trials.[6] Although a systematic review of placebo versus no treatment found little evidence for placebo effect,[7] one issue seems unquestionable: the placebo effect is present in clinical practice and in clinical trials — by whichever name we choose to call it or the nature of the phenomenon — and its amplitude may vary with drug treatment.[6] Therefore, the impact of placebo effects on statistical power in the context of PGx association studies needs to be evaluated and quantified.

Several factors have been shown to influence power estimations for association studies, such as: disease penetrance and prevalence; the net effect of the susceptibility locus; the frequency of the disease allele(s); the frequency of the marker allele(s); and the extent of linkage disequilibrium (LD) between disease alleles and marker alleles.[8,9] At present, there are no analytical derivations of power estimation that handle more realistic situations, such as complex dependencies between linked markers and the disease-causing allele frequency, recombination hot spots etc. Therefore, the strategy of choice is simulations. Long and Langley[10] pursued this strategy to quantify the power of complex trait association studies across a wide range of settings using a large number of simulations. They simulated genotypes based on the coalescent model,[11] phenotypes were randomised, with phenotype probability being conditioned on the causative single nucleotide polymorphism (SNP) genotypes, and association was evaluated using appropriate statistical tools. The study concluded that greater power was achieved by increasing the sample size than by increasing the number of polymorphisms, and that marker-based tests were more powerful than simple haplotype-based analyses.

PGx studies differ considerably from standard case-control association studies, however, as illustrated above and confirmed by our results; hence, it is important to quantify the statistical power of association studies in the context of PGx and to map the parameter space of such studies. Power estimation for PGx studies has been previously studied by Cardon *et al.*,[12] who used analytical formulae to study simplistic trial designs. They explored how different properties of SNPs, for example the frequency of the disease-causing alleles, might influence the required size and expected power of the clinical trial. Unfortunately, for PGx studies — as for complex trait associations — the frequencies of these phenotype-causing variants are unknown and their distribution is complex, motivating a simulation-based approach.

The goal of this study was to evaluate the power of PGx association studies under different scenarios, with special emphasis on the placebo effect. The setting was a drug clinical trial consisting of a double-blind, randomised controlled study, which included a placebo-treated control group and a drug-treated group. SNPs for a candidate gene region were then genotyped in these groups and tested for association with the response phenotype under the assumption of complete LD. Drug response was simplistically treated as a binary trait, and marker allele frequencies were then compared between responders (cases) and non-responders (controls), similar to a case-control design nested within a cohort.[13] Power was estimated by simulation, as in the study by Long and Langley,[10] and association was evaluated using a logistic regression model.[14,15] Since a considerable fraction of responders were expected to respond, due to the placebo phenocopy (an indistinguishable phenotype unrelated to the tested causative allele), we focused on the interaction between genotype and drug/placebo labelling. The model we propose assumes that specific genotypes have differential effects in the drug-treated group but not in the placebo-treated group.[12] Thus, the logistic regression term, which is expected to indicate true association, is the interaction term for genotype by drug. Various studies (eg Gauderman[16]) have calculated the required sample size for studies of gene–environment interactions, but the methods suggested are usually applicable to very specific designs and calculations are presented for specific sets of parameters and are therefore not directly applicable to the PGx context and the particular design of interest (randomised controlled study).

Power was estimated over a wide range of experimental design parameters: first and foremost, the number of individuals that participated in the clinical trial, the magnitude of the placebo effect and the penetrance of the response locus. We further examined direct (typing the causative allele itself) versus indirect (typing a tightly correlated SNP) tests and haplotype versus single marker frequency analyses. We also changed the ratio between the sizes of placebo- and drug-treated patient groups, the number of SNPs and the method for choosing those SNPs (either randomly or categorised in allele frequency bins).[9,17] Combined, our analyses provide a comprehensive examination of the parameter space for PGx study designs.

Materials and methods

For each setting of parameters, we evaluated power as the fraction of simulations, out of $R = 100$ or 1,000 (see below) repetitions, in which true association was detected, with an expected type I error of 5 per cent. Each of the R simulations was performed as outlined below:

- Generate genotype data
- Generate phenotype data
- For indirect tests, select SNPs for study
- Assess association between marker alleles/haplotypes and phenotype.

Parameters tested

We evaluated statistical power, as a function of the number (N) of individuals in the clinical trial ($N = 100$ to $N = 1{,}500$), under a range of different parameter settings:

- The frequency ($f_0 = 15$ per cent to $f_0 = 40$ per cent) of the placebo-response phenocopy. Importantly, this magnitude of the placebo effect is assumed to equal the penetrance (frequency of response) among homozygotes for the non-response allele.
- The size ratio between drug- and placebo-treated patient groups (either by suggesting a different study design — ie fixing the total number of patients — or by suggesting drug-only follow-up studies, fixing the number of placebo-treated individuals).
- The genotype relative risk (GRR) of the response polymorphism (2 to 4). GRR is defined as the ratio between the penetrance among homozygotes for the response allele (f_2) and homozygotes for the non-response allele (or placebo effect, f_0).[18]
- The number of SNPs examined ($M = 3$ or $M = 5$).
- The strategy for SNP selection (randomly or by frequency categories).

Generation of genotype data

The coalescent approach[11] was used to generate samples consisting of completely linked SNPs. A simple population genetic model involving only mutation and random genetic drift was assumed, without recombination within the small region considered. We simulated a fixed number of sites, using the *ms* software (see Hudson[19] for further details on haplotype generation). A single realisation of the coalescent process resulted in a set of haplotypes for 50 polymorphic sites. Sites were correlated, as expected by sites in complete LD. One of the sites was randomly chosen as the response site. The only requirement was that the frequency of its minor allele was more than 5 per cent. To further simplify the model, the ancestral allele was assigned as the aetiological allele. Haplotypes were then randomly paired to form genotypes.

Generation of phenotypic data

Patients were randomly assigned to the drug- or placebo–treated group with equal probability, or according to a fixed drug/placebo group size ratio. Patients assigned to the placebo group were randomly defined as responders or non-responders, with the probability of the former equal to the 'placebo effect'. Patients assigned to the drug group were randomly labelled responder/non-responder, with the probability of response determined by the penetrance of each genotype. For the non-response homozygotes, this probability was equal to the placebo effect. The penetrance of the heterozygote was set to the mean of the two homozygote penetrances, representing an additive mode of inheritance.

Strategy for SNP selection

$M = 3$ or $M = 5$ markers out of the 50 simulated markers in the candidate region were selected for genotyping. The number of SNPs per gene was limited to adhere to the budget constraints of the experimental design and, more importantly, availability: SNPs must be known (as if mined from public databases), technically typeable and polymorphic in the study population(s). The causative SNP was not explicitly excluded and could appear as one of the markers. Two strategies were tested for selecting the SNPs for genotyping:

- *Category approach*. In the presence of LD, adequate matching of allele frequencies at marker and trait loci determines if a marker site will be useful for detecting an association with the trait variant.[9,17] Following this principle, SNPs were classified into three or five distinct categories by their minor allele frequencies. One SNP from each category was then selected at random. If one category was empty of SNPs, we 'walked' along the chromosome until hitting a SNP with a frequency not already present in the selected set. The frequency categories were: 0.1–0.2, 0.2–0.35 and 0.35–0.5 for $M = 3$ markers; and 0.05–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4, 0.4–0.5 for $M = 5$ markers.
- *Random approach*. M different SNPs with minor allele frequencies greater than 10 per cent (for $M = 3$) or 5 per cent (for $M = 5$) were randomly chosen from the entire dataset. Two SNPs were allowed to have equal minor allele frequencies.

Detecting association between markers and drug response

Association was detected by a logistic regression model commonly used to analyse categorical data. We used the commercially available SAS statistical software.[20] In this analysis, the log odds of being a responder was regressed on the independent variables. The model contained two independent variables — a 'drug' indicator variable D (drug or placebo) and the genotype variable G (having three possible values: 0, 1 or 2) — and the interaction between them ($D^{\star}G$), namely

$$\text{Log odds} = \beta_0 + \beta_1 D + \beta_2 G + \beta_3 D^{\star} G,$$

where β_0 is the intercept and β_i ($i = 1$ to 3) is the change in log odds as a result of a unit increase in D, G, or $D^{\star}G$, respectively. Association was detected by a significant ($p < 0.05$) drug by genotype interaction effect. Intuitively, this is just a more general version of implementing an association test of responders versus non-responders in a drug-only experimental design, while accounting for the level of the placebo phenocopy, known from a separate, placebo-only design.

Two approaches were considered:

- A *'direct association'* approach, in which potential drug-response variants were tested one at a time. The suspected

causative SNP was therefore the only genotype considered in the logistic regression model. In this approach, $R = 1{,}000$ iterations were performed.

- An *'indirect association'* approach, in which several markers (three or five) were typed, hopefully turning out to be significantly correlated with the response locus. Genotypes of all of the SNPs were therefore considered in the logistic regression model, either marker by marker (testing each of the three or five SNPs with separate regression models and recording the highest statistic, as explained below) or as haplotypes. The individual contribution of each SNP varied, as expected between different random runs of the simulation process, and we focused on the overall significance of association. The significance of single-marker association was computed through a Monte Carlo permutation approach[21] and compared with haplotype analysis. For all indirect marker-based tests, which employed a Monte Carlo procedure[22] for power estimation, $R = 100$ was used, due to the computationally intensive nature of this analysis.

To assess the significance of single-marker association, we applied logistic regression analysis to each genotyped marker and recorded the highest statistic (Wald χ^2) for the drug by genotype interaction term. We randomly permuted the response labels and repeated the same analysis 500 times to obtain the distribution of the maximum χ^2 score under the null hypothesis of no association. The p value for a given simulation was estimated according to this distribution.

Haplotype analysis was more straightforward, since it did not require maximisation over many single marker scores. In this case, the logistic regression model included haplotypes and drug by haplotype terms, instead of the respective genotype terms. A haplotype variable assumes a value in $\{0,1,2\}$, denoting its copy number in the genotype of an individual. Haplotypes are assumed to be resolved by pedigrees or computation (eg Stephens *et al.*[23]). Note that the combination of complete LD and the selection of non-redundant SNPs implied that there are exactly $M + 1$ haplotypes. $R = 1{,}000$ simulations were run.

Type I error

Naturally, power should be compared when the false-positive rates are fixed to be the same across different methods. The statistical tests performed in these simulations were designed to hold the type I error at a constant rate of 5 per cent. To validate the rate of our type I error, simulations were run with GRR equal to 1 — ie f_2 was equal to the placebo effect. The proportion of false associations was then recorded for the different tests: direct analysis on the causative effect, the single-marker Monte Carlo permutation approach and the haplotype analysis for $N = 500$ and $N = 1{,}000$. The probability of detecting a false association was estimated when the placebo effect was 26 per cent (as in GRR = 3). The results of this validation benchmark are shown in Table 1. Note that the variance in false-positive rates for random SNPs seemed to be higher than that for haplotypes.

Comparison with predictions by existing tools

In order to compare the numbers obtained in this study with a scenario in which there was no placebo effect, power was calculated with the 'Genetic Power Calculator' (GPC) program,[24] for a 'classical' case-control study. The parameters were set as follows: GRR = 2, $f_2 = 0.4$, $f_0 = 0.2$, frequency of the response allele and marker allele = 0.7 (which is the mean frequency resulting from the coalescent simulation), complete LD, prevalence of response among drug-treated individuals = $0.34(0.7 \times 0.7 \times 0.4 + 2 \times 0.7 \times 0.3 \times 0.3 + 0.3 \times 0.3 \times 0.2)$, and a case:control ratio of 1.

Results

We first examined the power under the optimistic assumption of detecting direct association (ie the tested marker is the

Table 1. Estimated false-positive rates for the different statistical tests.

Number of persons	False-positive rates				
	Direct association		Indirect association		
			Single marker		Haplotype
			Categories	Random	
500	0.045	$M = 3$	0.06	0.06	0.051
		$M = 5$	0.03	0.08	0.039
1,000	0.042	$M = 3$	0.04	0.05	0.046
		$M = 5$	0.06	0.01	0.04

M is number of markers typed.

causative SNP). In Figure 1a, power is plotted as a function of the total number of persons participating in the clinical trial (half placebo-treated and half drug-treated) for different penetrance scenarios. Even for the best penetrance-scenario examined (GRR = 3 and placebo effect $f_0 = 26.6$ per cent), more than 500 individuals are required to be included in the clinical trial to reach the standard level of 80 per cent power. This is in sharp contrast to the predictions of the GPC,[24] which are an order of magnitude smaller than the worst penetrance scenario examined (in Figure 1a, compare GPC [plotted in dashed curve] and GRR with $f_2 = 0.4$, $f_0 = 0.2$). Another observation is that the power curves are sorted according to the penetrance of the response genotype, f_2. This may be expected, given that GRR $= f_2/f_0$ and that the prevalence of response is a function of f_2, and f_0. To better evaluate the relative impact of the penetrances f_2 and f_0 on power, in Figure 1b we plotted the power as a function of these parameters for a fixed number ($N = 1{,}000$) of persons per trial. Fixing each of these penetrance parameters reveals that the power, across its dynamic range, is almost a linear function of the other penetrance. We can observe that for a given value of f_2, power decreases approximately linearly as f_0 increases. Moreover, for a given value of f_0, power increases approximately linearly as function of f_2 at most of the power ranges. In addition, for a given GRR ratio, power is considerably affected by the value of f_2. Thus, considering the parameter space defined in our simulations, power for $f_2 = 0.4$, 0.6 or 0.8, and GRR = 2, is 0.39, 0.644 and 0.881, respectively.

Figure 2 presents the effect of different drug/placebo group size ratios on power for the best penetrance scenario in Figure 1a (GRR = 3, $f_2 = 0.8$, $f_0 = 0.266$). Figure 2a refers to the scenario where a first clinical trial including drug- and placebo-treated groups has been completed and, in order to enlarge the sample size, drug-only follow-up studies are included in subsequent analyses. We therefore used a fixed number of placebo-treated individuals and increased the size of

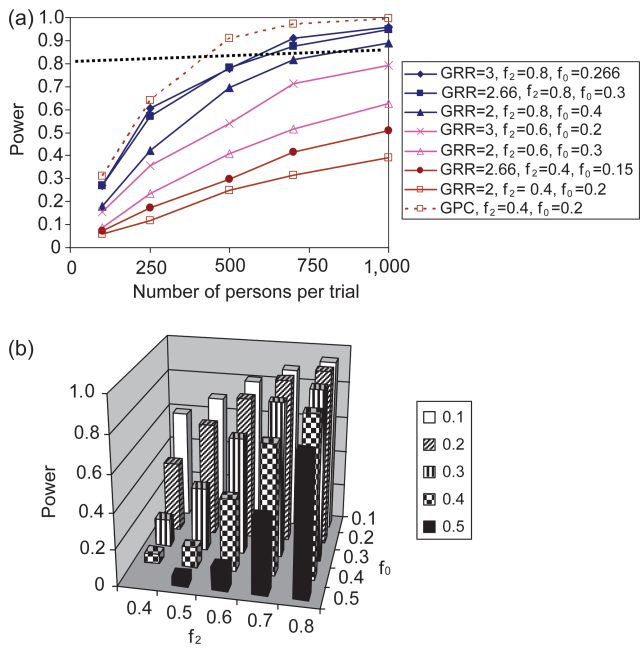

Figure 1. (a) The effect of genotype relative risk (GRR) ratio and placebo effect (f_0) on the power of detecting an association between the causative single nucleotide polymorphism (SNP) and the response to the drug. Power estimates are plotted against the number of persons per trial for different values of f_2 (penetrance among homozygotes for the response allele), f_0 and GRR values. GPC refers to the Genetic Power Calculator program, which does not take into account the placebo effect. The horizontal broken line denotes the 80 per cent power threshold. (b) The relative effect of f_2 and f_0 on the power of detecting an association between the causative SNP and the response to the drug. The number of persons per trial equals 1,000. Results are based on 1,000 simulations.

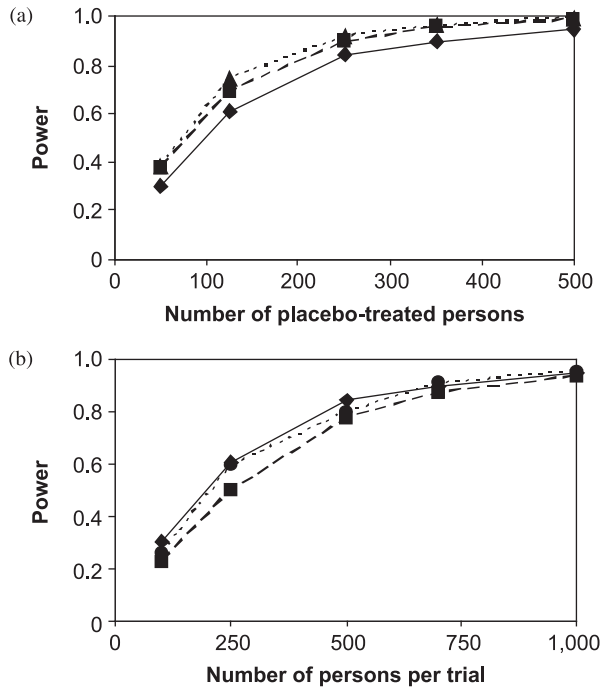

Figure 2. The effect of increasing the ratio between the drug and placebo group sizes on the power of detecting an association between the causative single nucleotide polymorphism and the response to the drug. Estimated power is plotted against the number of persons who received placebo for the different drug to placebo group size ratios: 1:1 –◆–; 2:1 –■–; 4:1 --▲--. The results are based on 1,000 simulations. Genotype relative risk = 3, $f_2 = 0.8$, $f_0 = 0.266$. The results are presented as a function of: (a) the number of placebo-treated individuals (mimicking a drug-only follow-up); (b) the number of individuals in the entire trial.

the cohort of drug-treated patients. Plots of power versus study size for different ratios (1:1, 2:1 or 4:1) between placebo- and drug-treated group sizes are shown. Worst and intermediate penetrance scenarios in Figure 1a were also analysed (data not shown). Increasing the number of drug-treated patients improved power only minimally (usually 10–25 per cent for the first doubling and an additional <15 per cent for the second). Improvement was largest for the less powered scenario (data not shown). To evaluate the best design for a PGx association study when the number of patients is limited, we calculated power for 1:1, 2:1 and 1:2 drug-/placebo-treated group size ratios for the same GRR/f_2/f_0 scenarios as in Figure 2a, but this time fixing the total number of patients participating in the clinical trial. While the best ratio seems to be 1:1, and the worst 2:1, differences are small and often statistically insignificant (Figure 2b).

We next evaluated power for the indirect approach — ie the tested marker is distinct from the response SNP (Figures 3–5). The power curve for analysis, including the causative SNP, is also presented for comparison. In Figure 3, we compared power for two different strategies for selecting the markers to be genotyped, either randomly or by categories (see Methods section for details), examining three penetrance scenarios and two options for the number of markers typed ($M = 3$ or $M = 5$). Only for the most empowered setting (Figure 3f) did the 'categories strategy' show a consistent advantage over the 'random strategy'.

Comparing power obtained for the different number M of markers typed on the same simulated datasets yielded similar plots, with enhanced power for $M = 5$ over $M = 3$ (Figure 4). This improvement is large for larger study sizes and it is significant (see grey-shaded patches in Figure 4), even for the modest number of performed simulations when the study size is increased.

We used the same datasets (categories strategy) to compare the relative power of haplotype versus single marker analysis (Figure 5). Perhaps surprisingly, straightforward haplotype analysis does not seem to have an advantage over single marker analysis (which seems superior in the scenarios examined in Figures 5b and 5f). Furthermore, neither of the power plots for graphs 5a–f indicate statistically significant differences between these analytical approaches.

Discussion

We have shown that the attributes characteristic of a clinical trial, particularly the magnitude of the placebo effect, have unexpected implications on the statistical power of PGx association studies. Our simulation results stand in sharp contrast to the over-optimistic predictions of tools designed primarily for case-control disease association studies[24] and highlight the marked impact that a substantial placebo effect can have on reducing study power. In the absence of analytical tools specifically tailored to calculate power in the PGx context, where gene–environment interactions are integrated our results can only be compared with tools designed for classical disease association studies. The simulation study presented here shows that even under the most favourable scenario — involving high penetrance conditions — reliable association (80 per cent power) between SNPs in a candidate gene or region and the response to a drug requires the recruitment of an 'optimal number' — $N \approx 500$ patients — in a clinical trial, given that the causative SNP is genotyped, and $N \approx 800$ patients when five perfectly linked markers are genotyped (Figure 4). Despite the fact that for some results regarding the indirect association approach the standard errors are still large (due to limited number of simulations performed), a general trend is nevertheless visible. It is hence crucial to take the marked impact of the placebo effect on power into consideration in PGx studies. Our empirical approach allows exploration of a complex array of practical issues of study design, in contrast to previous, theoretical, simplistic studies.[12] Therefore, the results presented here are meant to guide the optimal integration of genotype data into ongoing clinical trials and to define the size of such a trial required for a PGx study.

In practice, once a beneficial effect of a new treatment is clearly demonstrated, patients on placebo treatment are shifted to real therapeutic regimens. Hence, the total size of a given placebo-treated cohort will often remain limited, while the number of drug-treated patients will potentially significantly increase. We report in this study that the optimal study design in the presence of a placebo effect under the models examined comprises an equal number of drug- and placebo-treated patients, as is usually the case in Phase III clinical trials. Adding more drug-treated patients, even four times as many, increases power only mildly. This is in sharp contrast to the more classical case-control studies aimed at the elucidation of the aetiology of common diseases, where the number of affected cases is the limiting variable and where significant gains in power could be obtained by increasing the size of the control group.[9] We speculate that the rationale for this differential impact of relative cohort sizes is that in PGx it is essential to evaluate the penetrance for the non-causative genotype (f_0), which is negligible in disease susceptibility, and therefore the number of placebo-treated individuals becomes a tighter bottleneck.

A further potential improvement for the study design is an educated selection of markers. Ideally, markers need to be chosen in such a manner as to improve the chances of matching the causative allele frequency.[8,9] Yet, the latter is unknown (ie whether common as proposed under the 'common-disease, common-variant hypothesis'[25]) or less frequent, as also advocated.[26] Even though detailed haplotype maps[27] are well underway, which may eventually allow SNP selection based on phylogenetic analysis[28] or haplotype blocks,[29] until such data are understood, one is still restricted to choosing markers from a modest set of validated SNPs, often with allele frequencies being the only additional data available. In this study, we spread marker frequencies over the

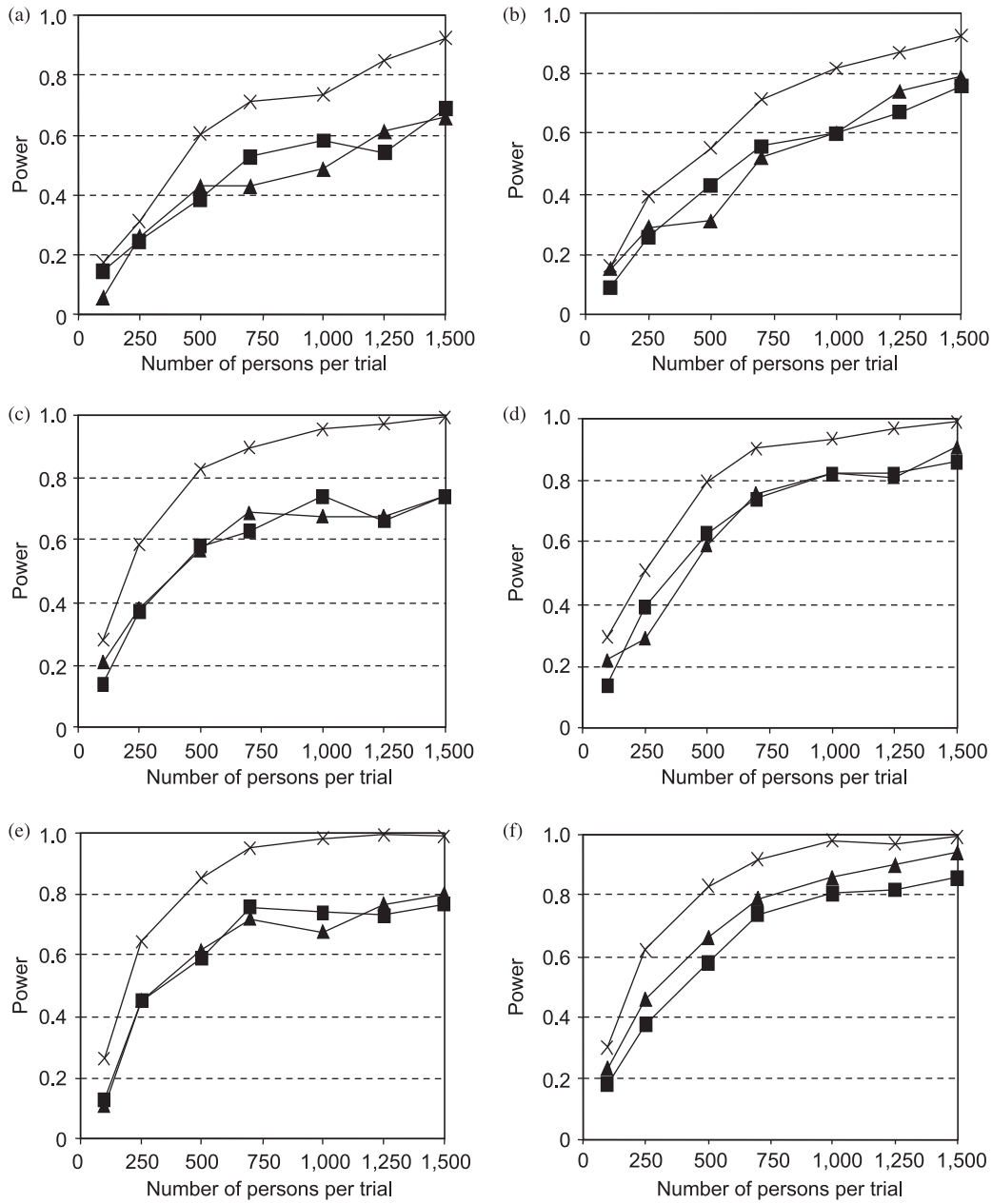

Figure 3. The effect of strategy for selecting markers (*M*, at random or categorised) on the power of detecting an association between the markers and the response to the drug. The results are based on 100 simulations. Estimated power is shown for the causative single nucleotide polymorphism (SNP) (–**x**–), randomly selected SNPs (–■–) and a categories-based strategy (–▲–). (a) $M = 3$, genotype relative risk (GRR) $= 3$, $f_2 = 0.6$, $f_0 = 0.2$; (b) $M = 5$, GRR $= 3$, $f_2 = 0.6$, $f_0 = 0.2$; (c) $M = 5$, GRR $= 3$, $f_2 = 0.8$, $f_0 = 0.266$; (d) $M = 5$, GRR $= 4$, $f_2 = 0.8$, $f_0 = 0.2$; (e) $M = 3$, GRR $= 4$, $f_2 = 0.8$, $f_0 = 0.2$; (f) $M = 5$, GRR $= 4$, $f_2 = 0.8$, $f_0 = 0.2$.

possible range of informative alleles (>5 per cent or >10 per cent). We compared this strategy with that of choosing markers randomly. Surprisingly, little difference in power is reported, if at all. One possible explanation might be that redundant markers are not the major source of power loss when only a small set of markers is used, as these SNPs are likely to fall in different allele frequency categories by chance. Yet our results suggest that power is greatly increased if five markers ($M = 5$) are typed instead of three ($M = 3$) (Figure 4), as with case-control association studies. This is likely to stem from the increased chances, as *M* gets larger, of hitting a marker allele which is in phase with the response allele. Since

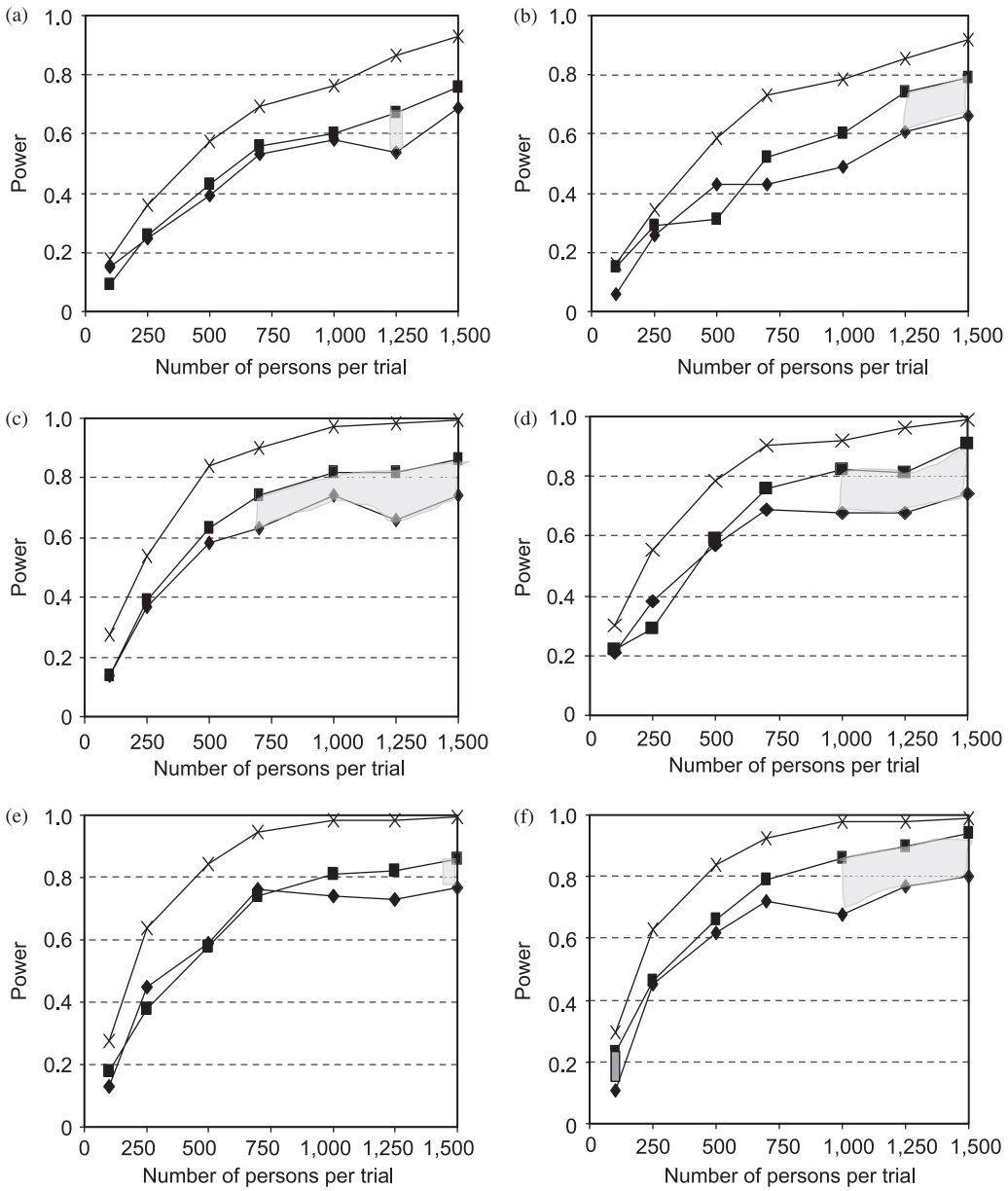

Figure 4. The effect of the number (*M*) of markers typed (three or five) on the power of detecting an association between the markers and the response to the drug. The results are based on 100 simulations. The estimated power is shown for $M = 3$ (–◆–), $M = 5$ (–■–) and the causative single nucleotide polymorphism (–**x**–). The shaded regions denote designs for which the difference between $M = 3$ and $M = 5$ is significant. (a) Random selection, genotype relative risk (GRR) = 3, $f_2 = 0.6$, $f_0 = 0.2$; (b) Categories strategy, GRR = 3, $f_2 = 0.6$, $f_0 = 0.2$; (c) Random selection, GRR = 3, $f_2 = 0.8$, $f_0 = 0.266$; (d) Categories strategy, GRR = 3, $f_2 = 0.8$, $f_0 = 0.266$; (e) Random selection, GRR = 4, $f_2 = 0.8$, $f_0 = 0.2$; (f) Categories strategy, GRR = 4, $f_2 = 0.8$, $f_0 = 0.2$.

the number of individuals participating in a clinical trial is limited, increasing the number of genotyped markers may be the strategy of choice, and the only feature controlled by study designers, for improving the power of a PGx association study.

In this study, we also considered the option of improving power by a higher-level analysis of the genotypic data. Our simulations extend earlier results in a complex-trait context[10,29] to the PGx framework, regarding similarity of power in analysis based on haplotypes versus single markers. More sophisticated analysis of haplotypes, exploiting their cladistic structures, may, however, be more advantageous in PGx than in other areas,[30,31] yet the impacts of a departure from the infinite site model (an assumption implicit in our coalescent simulation) and of homoplasy remain to be

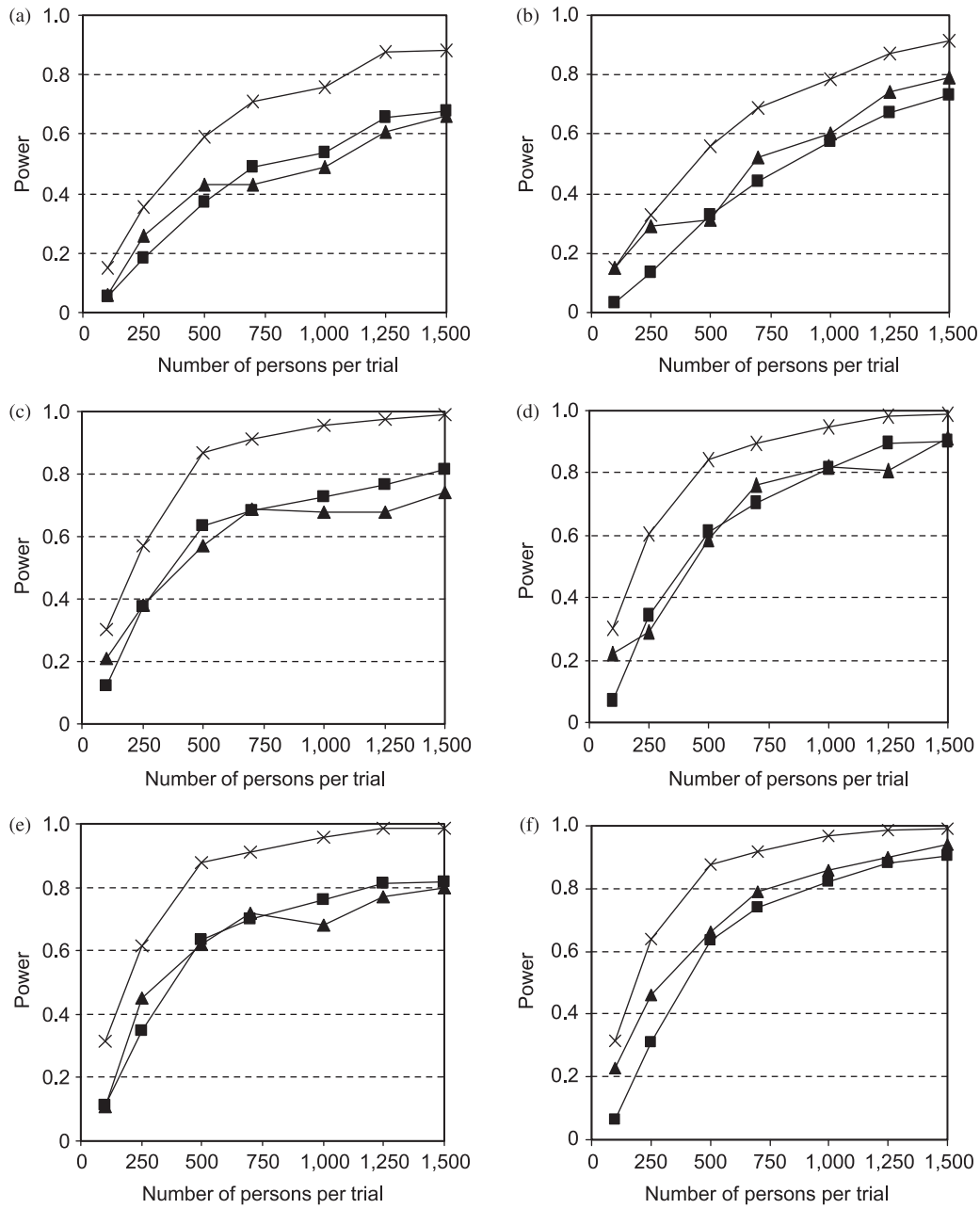

Figure 5. The effect of haplotype analysis versus one single nucleotide polymorphism (SNP) analysis on the power of detecting an association between the markers and the response to the drug. The results are based on 100 simulations for marker analysis and 1,000 simulations for haplotype analysis. Markers (*M*) were chosen according to frequency categories. The estimated power is shown for the causative SNP (–**x**–), haplotype analysis (–■–) and single marker (–▲–) analysis. (a) $M = 3$, genotype relative risk (GRR) $= 3$, $f_2 = 0.6$, $f_0 = 0.2$; (b) $M = 5$, GRR $= 3$, $f_2 = 0.6$, $f_0 = 0.2$; (c) $M = 5$, GRR $= 3$, $f_2 = 0.8$, $f_0 = 0.266$; (d) $M = 5$, GRR $= 4$, $f_2 = 0.8$, $f_0 = 0.2$; (e) $M = 3$, GRR $= 4$, $f_2 = 0.8$, $f_0 = 0.2$; (f) $M = 5$, GRR $= 4$, $f_2 = 0.8$, $f_0 = 0.2$.

calibrated. These results place another pin on the map of the literature on haplotype versus single marker analyses, each method having its own advantages.[10,29,32,33]

The frequency of the response allele is an important determinant of the power of association studies.[8,29] Since this aspect of association studies has been extensively analysed, however, we avoid handling this issue, relying instead on existing analysis.

Simulation assumptions in this study consider a very basic genetic model: an equilibrium population with only mutation and random genetic drift modifying a non-recombinant

haplotype block containing the candidate gene under study. Real life is far more complex. Nonetheless, this model is already sufficient to indicate the general trends of the factors that may confound PGx studies. While this simple model does not accurately reflect samples drawn from human populations, we consider it preferable to more assumptive, but often still controversial, models. Incorporating other factors, such as recombination, gene conversion, recurrent mutations or demographic expansion, into the coalescent model is likely to deteriorate the power estimated in the present study. It should be noted that we make implicit assumptions in the manner in which simulations are laid out. First, the response allele is assumed to be the ancestral, usually more common, one. This assumption is rationalised by our focus on drugs that, by default, do evoke a response, by contrast with long-shot treatments whose success is the exception and which require separate analysis. Furthermore, the range of minor allele frequencies that are examined in this work may bias our findings. The simulation parameters analysed implicitly focus this work at more common SNPs, more akin to the common-disease, common-variant scenario. Other excluded factors relevant specifically to a PGx power study — such as multiple drug doses, quantitative or categorical outcomes instead of a binary response, different models for placebo effect, allelic heterogeneity, epistatic interactions and genotyping errors — all motivate further research. Lastly, studies of adverse drug effects, which are not examined in the current study, may require further research involving this particular design.

The interest of large pharmaceutical companies in PGx studies, the strong possibility that new drugs will be required to be evaluated for PGx by the Food and Drugs Administration and the public demand for more personalised medicines is likely to increase the number of PGx studies in the near future. To increase the likelihood of obtaining significant results, studies need to be designed to take into consideration the parameters that affect power estimation. The present study implies that simple transpositions of conventional case-control models and power evaluations to PGx are not straightforward and require separate consideration. While statistical power in PGx is affected by some parameters, as with disease susceptibility studies, the particularities of a study design that is based on a clinical trial change the set of controllable parameters and transform the landscape of success probabilities. The follow-ups suggested above are expected to further refine the outline characteristics of statistical power in PGx studies of drug response.

Acknowledgments

We thank Alan Templeton, Edna Schechtman and Doron Lancet for helpful comments on this work. This work was supported by funds provided by the 'Magneton Program' — a combined project with Teva Pharmaceuticals Ltd and the Office of the Chief Scientist, Ministry of Industry and Trade, Israel. I.P. is a recipient of the ESHKOL fellowship by the Israeli Ministry of Science and Technology. J.S. Beckmann holds the Hermann-Mayer chair and was supported by the Henry S. and Anne S. Reich Research Fund. Finally, we thank the anonymous reviewers for their insightful comments.

References

1. Roses, A.D. (2000), 'Pharmacogenetics and the practice of medicine', *Nature* Vol. 405, pp. 857–865.
2. Pirmohamed, M. and Park, B.K. (2001), 'Genetic susceptibility to adverse drug reactions', *Trends Pharmacol. Sci.* Vol. 22, pp. 298–305.
3. Lurcott, G. (1998), 'The effects of the genetic absence and inhibition of CYP2D6 on the metabolism of codeine and its derivatives, hydrocodone and oxycodone', *Anesth. Prog.* Vol. 45, pp. 154–156.
4. Arranz, M.J., Munro, J., Birkett, J. *et al.* (2000), 'Pharmacogenetic prediction of clozapine response', *Lancet* Vol. 355, pp. 1615–1616.
5. Morton, V. and Torgerson, D.J. (2003), 'Effect of regression to the mean on decision making in health care', *BMJ.* Vol. 326, pp. 1083–1084.
6. Spiegel, D., Kraemer, H. and Carlson, R.W. (2001), 'Is the placebo powerless?', *N. Engl. J. Med.* Vol. 345, pp. 1276; author reply pp. 1278–1279.
7. Hrobjartsson, A. and Gotzsche, P.C. (2001), 'Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment', *N. Engl. J. Med.* Vol. 344, pp. 1594–1602.
8. Zondervan, K.T. and Cardon, L.R. (2004), 'The complex interplay among factors that influence allelic association', *Nat. Rev. Genet.* Vol. 5, pp. 89–100.
9. McGinnis, R., Shifman, S. and Darvasi, A. (2002), 'Power and efficiency of the TDT and case-control design for association scans', *Behav. Genet.* Vol. 32, pp. 135–144.
10. Long, A.D. and Langley, C.H. (1999), 'The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits', *Genome Res.* Vol. 9, pp. 720–731.
11. Hudson, R.R. (1983), 'Properties of a neutral allele model with intergenic recombination', *Theor. Popul. Biol.* Vol. 23, pp. 183–201.
12. Cardon, L.R., Idury, R.M., Harris, T.J. (2000), 'Testing drug response in the presence of genetic information: Sampling issues for clinical trials', *Pharmacogenetics* Vol. 10, pp. 503–510.
13. Essebag, V., Genest, Jr., J., Suissa, S. *et al.* (2003), 'The nested case-control study in cardiology', *Am. Heart J.* Vol. 146, pp. 581–590.
14. Ott, J. and Hoh, J. (2001), 'Statistical multilocus methods for disequilibrium analysis in complex traits', *Hum. Mutat.* Vol. 17, pp. 285–288.
15. Zee, R.Y., Hoh, J., Cheng, S. *et al.* (2002), 'Multi-locus interactions predict risk for post-PTCA restenosis: An approach to the genetic analysis of common complex disease', *Pharmacogenomics J.* Vol. 2, pp. 197–201.
16. Gauderman, W.J. (2002), 'Sample size requirements for matched case-control studies of gene–environment interaction', *Stat. Med.* Vol. 21, pp. 35–50.
17. Garner, C. and Slatkin, M. (2003), 'On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci', *Genet. Epidemiol.* Vol. 24, pp. 57–67.
18. Clayton, D. (2001), 'Population association', in Balding, D.J., Bishop, M., Cannings, C. (eds), *Handbook of Statistical Genetics*, John Wiley and Sons, NewYork, NY, pp. 519–540.
19. Hudson, R.R. (2002), 'Generating samples under a Wright–Fisher neutral model of genetic variation', *Bioinformatics* Vol. 18, pp. 337–338.
20. Anon. (1999), 'The logistic procedure' chapter 39 in *SAS/STAT User's Guide*, Version 8, SAS Publishing, SAS Institute Inc Cary, NC, pp. 1903–2044.
21. Churchill, G.A. and Doerge, R.W. (1994), 'Empirical threshold values for quantitative trait mapping', *Genetics* Vol. 138, pp. 963–971.
22. McIntyre, L.M., Martin, E.R., Simonsen, K.L. *et al.* (2000), 'Circumventing multiple testing: A multilocus Monte Carlo approach to testing for association', *Genet. Epidemiol.* Vol. 19, pp. 18–29.
23. Stephens, M., Smith, N.J. and Donnelly, P. (2001), 'A new statistical method for haplotype reconstruction from population data', *Am. J. Hum. Genet.* Vol. 68, pp. 978–989.
24. Purcell, S., Cherny, S.S. and Sham, P.C. (2003), 'Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits', *Bioinformatics* Vol. 19, pp. 149–150.

25. Lander, E.S. (1996), 'The new genomics: Global views of biology', *Science* Vol. 274, pp. 536–539.
26. Pritchard, J.K. and Cox, N.J. (2002), 'The allelic architecture of human disease genes: Common disease-common variant . . . or not?', *Hum. Mol. Genet.* Vol. 11, pp. 2417–2423.
27. The International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
28. Templeton, A.R., Weiss, K.M., Nickerson, D.A. *et al.* (2000), 'Cladistic structure within the human lipoprotein lipase gene and its implications for phenotypic association studies', *Genetics* Vol. 156, pp. 1259–1275.
29. Barrett, J.C., Fry, B., Maller, J. *et al.* (2005), 'Haploview: Analysis and visualization of LD and haplotype maps', *Bioinformatics* Vol. 21, pp. 263–265.
30. Kaplan, N. and Morris, R. (2001), 'Issues concerning association studies for fine mapping a susceptibility gene for a complex disease', *Genet. Epidemiol.* Vol. 20, pp. 432–457.
31. Templeton, A.R., Boerwinkle, E. and Sung, C.F. (1987), 'A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*', *Genetics* Vol. 117, pp. 343–351.
32. Seltman, H., Roeder, K. and Devlin, B. (2003), 'Evolutionary-based association analysis using haplotype data', *Genet. Epidemiol.* Vol. 25, pp. 48–58.
33. Akey, J., Jin, L. and Xiong, M. (2001), 'Haplotypes vs single marker linkage disequilibrium tests: What do we gain?', *Eur. J. Hum. Genet.* Vol. 9, pp. 291–300.