

# “Forward Genetics” as a Method to Maximize Power and Cost-Efficiency in Studies of Human Complex Traits

M. P. M. Boks · E. M. Derks · C. V. Dolan ·  
R. S. Kahn · R. A. Ophoff

Received: 29 June 2009 / Accepted: 17 February 2010 / Published online: 16 March 2010  
© The Author(s) 2010. This article is published with open access at Springerlink.com

**Abstract** There is increasing interest in methods to disentangle the relationship between genotype and (endo)phenotypes in human complex traits. We present a population-based method of increasing the power and cost-efficiency of studies by selecting random individuals with a particular genotype and then assessing the accompanying quantitative phenotypes. Using statistical derivations, power- and cost graphs we show that such a “forward genetics” approach can lead to a marked reduction in sample size and costs. This approach is particularly apt for implementing in epidemiological studies for which DNA is already available but the phenotyping costs are high.

**Keywords** Power · Forward genetics · Complex traits · Selective sampling · Phenomics

## Introduction

Several genome-wide association studies have recently identified novel susceptibility loci for medical conditions such as diabetes mellitus and schizophrenia (Barret et al. 2009; Stefansson et al. 2009). This has increased the need to investigate the phenotypic differences that are conferred by such quantitative trait loci. However, due to the small to modest contributions of single loci to most complex traits, such phenotypic differences are hard to detect. The problem of small effect sizes of specific alleles or haplotypes is compounded by the complexity of many of the traits of interest. Most candidate genes have been identified in subjects with a disorder that incorporates a broad variety of symptoms. The co-occurrence of several symptoms at once can be the result of pleiotropic effects of a single variant, but may also be due to the underlying abnormality. Many disorders coincide with physical, emotional, and social abnormalities; for example, depression is associated with

---

Edited by Pak Sham.

---

**Electronic supplementary material** The online version of this article (doi:10.1007/s10519-010-9348-y) contains supplementary material, which is available to authorized users.

---

M. P. M. Boks (✉)  
Rudolf Magnus Institute of Neuroscience, Department  
of Psychiatry, B01.206, University Medical Centre Utrecht,  
PO Box 85500, 3508 GA Utrecht, The Netherlands  
e-mail: mboks@umcutrecht.nl

M. P. M. Boks  
Julius Center for Health Sciences and Primary Care,  
University Medical Centre Utrecht, Utrecht, The Netherlands

E. M. Derks · R. S. Kahn  
Rudolf Magnus Institute of Neuroscience, Department  
of Psychiatry, University Medical Centre Utrecht,  
PO Box 85500, 3508 GA Utrecht, The Netherlands

C. V. Dolan  
Department of Psychology, Faculty of Social and Behavioral  
Sciences, University of Amsterdam, Amsterdam,  
The Netherlands

R. A. Ophoff  
Department of Medical Genetics and Rudolf Magnus Institute  
of Neuroscience, University Medical Center Utrecht, Utrecht,  
The Netherlands

R. A. Ophoff  
UCLA Center for Neurobehavioral Genetics, Semel Institute  
for Neuroscience and Human Behavior, Los Angeles, CA, USA

cognitive problems, cardiovascular risks and social problems, among others. These concomitant phenomena are likely to influence the expression of the original trait, and may well obscure the initial relationship between a candidate gene and a trait. As a consequence, the particular symptoms or abnormalities associated with these genes remain unclear.

We, therefore, propose reversing the process: instead of selecting a trait and examining its relationship with the underlying genes, we will select genetic variation and examine the accompanying trait. Testing the influence of a particular gene on phenotypes is a common approach in both animal and molecular research, where the influence of genetic variation is often studied by inbreeding the genetic variant by creating a knockout mouse, or by transposing the variant of interest into cell cultures or organisms by means of a vector. The statistical power of an association test for a candidate gene depends on the distribution of genotypes in the test population. A maximum statistical power for a given number of phenotyped individuals is obtained when the test population consists of equal numbers of alternative homozygotes at the candidate gene (as is the case when the two alleles at the candidate gene are of equal frequency). However, since allele frequencies at the candidate gene locus are generally far from equal, the distribution of informative alleles in the population as a whole is generally far from this optimal distribution. Thus, depending on the relative costs of determining phenotypes compared to the cost of genotypes, it may be more effective to genotype a large sample population and then choose a set of individuals with an optimal distribution of genotypes for further phenotyping. Here we provide information on the statistical power under different genotype sampling strategies, as a function of explained variance, dominance and allele frequency at the candidate gene, and on phenotype/genotype cost ratios.

Selecting subjects from the general population based on homozygosity for a candidate gene instead of subjects with an apparent disorder has two major advantages. (1) It means the investigation of the relationship with genotype is unbiased by selection for severity of disease, and we therefore avoid bias as a result of secondary symptoms. (2) This approach facilitates the estimation of the effects of single variants in relative isolation, because the selection is not based on phenotype. As a consequence, there is no selection for the presence of additional risk variants for that particular phenotype although it will shift the distribution of the phenotype.

The value of such a “forward genetics” approach is seen in the increase in statistical power and its cost-effectiveness. As already pointed out, the increase in power is due to the selection of the most informative subjects. Other

strategies, such as the extreme discordant and concordant design (Risch and Zhang 1995), in essence do the same by selecting extreme phenotypes. With the ever reducing costs of genotyping, our strategy only has merit if the cost of obtaining phenotype information is high. In studies of complex and quantitative phenotypes, such as those that apply costly neuroimaging, this approach can be particularly advantageous.

We investigated the sample size requirement of this approach and the cost-effectiveness under different scenarios.

## Methods

The sample size requirements depend on the proportion of variance explained by the genetic effect, allele frequencies, and the genetic model (Falconer and Mackay 1996). In our description of the genetic model, we follow Falconer’s notation, in which the mean genotypic values of the  $A_1A_1$ , the  $A_1A_2$ , and  $A_2A_2$  genotypes are denoted as  $+a$ ,  $d$ , and  $-a$ , respectively. The total variance of a phenotype ( $V_P$ ) can be broken down into the variance due to one particular locus ( $V_G$ ), and the remaining genotypic and environmental variance  $V_R$ .  $V_G$  can be further divided into the additive genetic ( $V_A$ ) and the dominant genetic variance components ( $V_D$ ), such that:

$$V_G = V_A + V_D = 2pq[a + d(q - p)]^2 + [2pqd]^2, \quad (1)$$

where  $p$  and  $q$  denote the allele frequencies of the wildtype and risk alleles, respectively. Rearranging formula (1) provides an expression for the genotypic value of the  $A_1A_1$  genotype as

$$a = \left[ V_G / \left( (2pq(1 + d/a(q - p)))^2 + (2pq)^2(d/a)^2 \right) \right]^{0.5} \quad (2)$$

Based on the chosen values of  $p$ ,  $q$ ,  $d/a$ ,  $V_G$ , the total number of selected individuals ( $N$ ), and the selected number of individuals from each genotype group, we can derive the statistical power from a calculation of the  $F$  statistic. In order to calculate the  $F$  statistic, we first calculated the total mean, the within-group means, the within-group sum of squares ( $ss_{\text{within}}$ ) and the between-group sum of squares ( $ss_{\text{between}}$ ). The within-group means of the three genotype groups do not change as a result of selection and are therefore equal to  $+a$ ,  $d$ , and  $-a$ , for the  $A_1A_1$ , the  $A_1A_2$ , and  $A_2A_2$  genotypes. Denoting  $pr[A_1A_1]$ ,  $pr[A_1A_2]$ , and  $pr[A_2A_2]$  as the proportions of the three genotype groups after selection, the total mean,  $ss_{\text{between}}$ , and  $ss_{\text{within}}$  were calculated as

$$\begin{aligned} \text{Total mean} &= \text{Mean}_t = \text{pr}[A_1A_1] * a + \text{pr}[A_1A_2] * d \\ &\quad - \text{pr}[A_2A_2] * a \\ \text{SS}_{\text{within}} &= V_R * N \\ \text{SS}_{\text{between}} &= \text{pr}[A_1A_1] * N * (a - \text{mean}_t) \\ &\quad + \text{pr}[A_1A_2] * N * (d - \text{mean}_t) \\ &\quad + \text{pr}[A_2A_2] * N * (\text{mean}_t - a) \end{aligned} \quad (3)$$

The F statistic is defined as

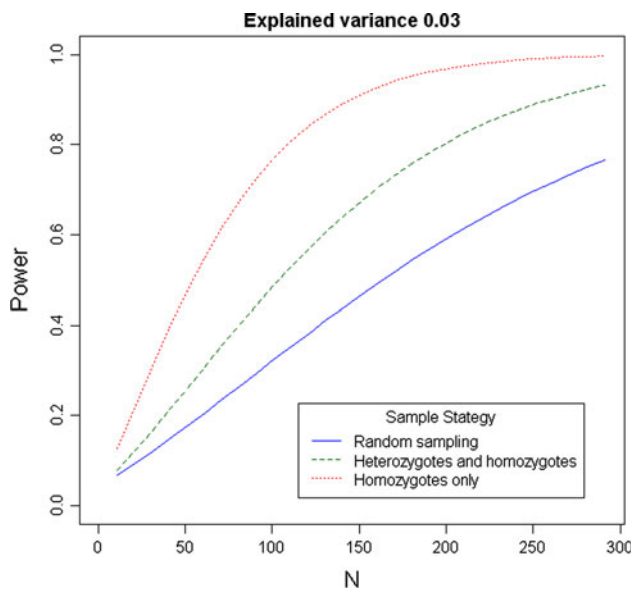
$$F = [\text{SS}_{\text{between}} / (k - 1)] / [\text{SS}_{\text{within}} / (N - k)] \quad (4)$$

with  $k$  denoting the number of groups, i.e.,  $k = 3$  (homozygotes and heterozygotes) or 2 (homozygotes only).

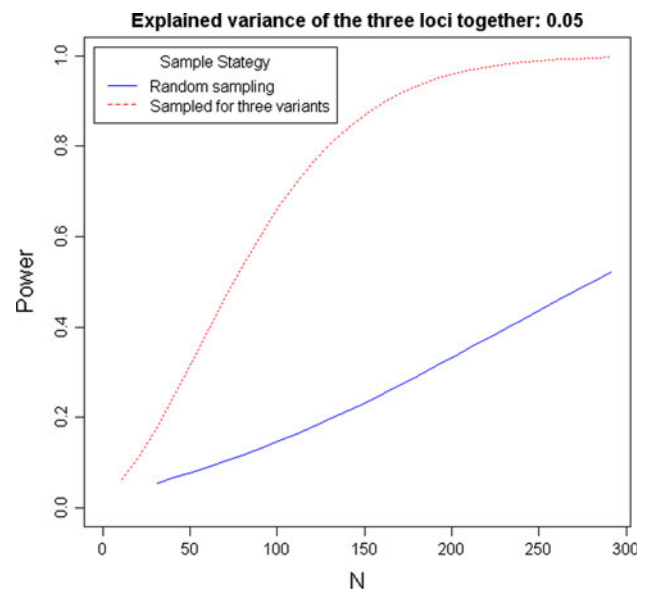
Finally, we calculated the non-centrality parameter (NCP), i.e.,  $F * (N - k)$ , and the statistical power based on this parameter and a chosen alpha. We compared the NCP calculated with that provided by the genetic power calculator (Purcell et al. 2003). The application can be extended for scenarios where more genes are considered (see supplementary R script). We performed illustrative power and costs analyses under several scenarios, including one in which we selected three genes and subjects were available that were homozygous for two of the rare homozygous variants and heterozygous for the third variant compared to subjects that were homozygous for all three common variants. The calculations were carried out in R (R Development Core Team 2005).

## Results

Figure 1 shows the power with a type I error rate of 0.05 for different sampling strategies in a fully additive model,

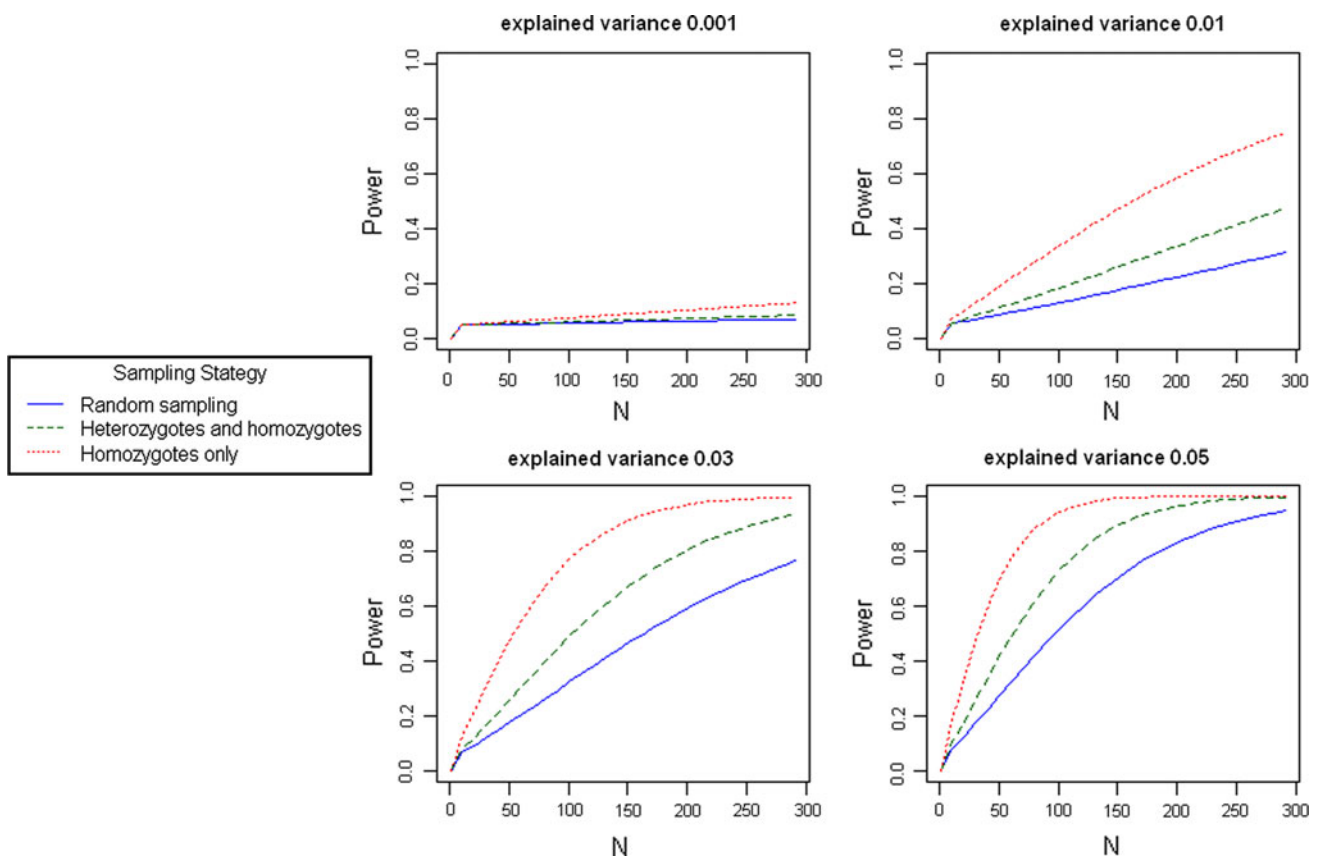


**Fig. 1** Power achieved under different sampling strategies

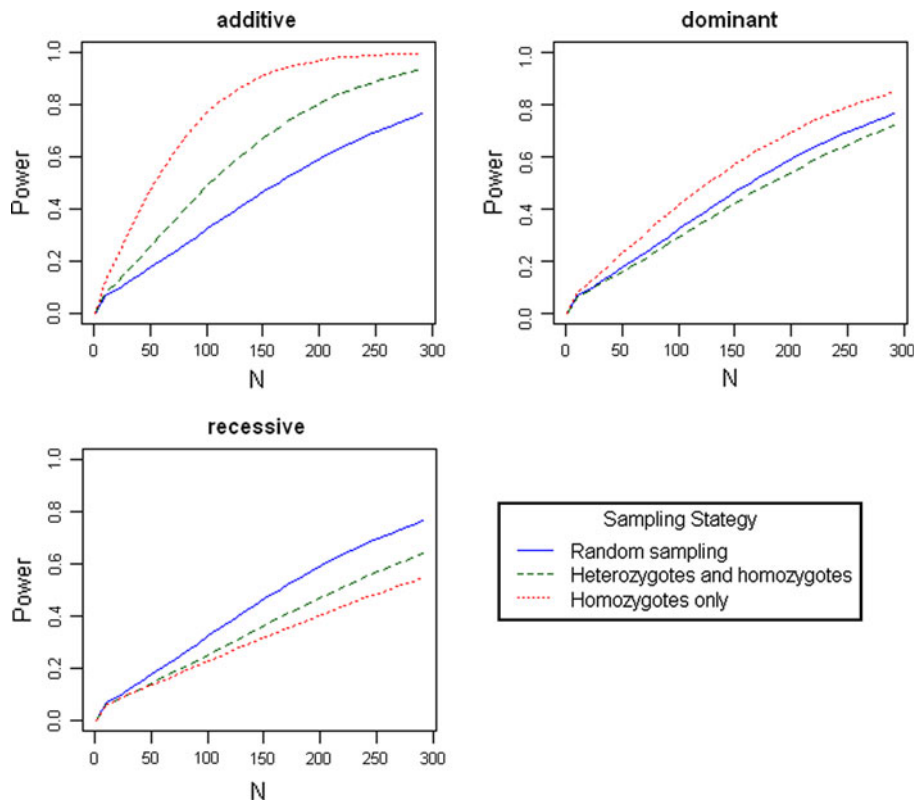


**Fig. 2** Power for a sampling strategy where three genes were selected and subjects were available that were homozygous for two of the rare homozygous variants and heterozygous for the third variant compared to subjects that were homozygous for all three common variants

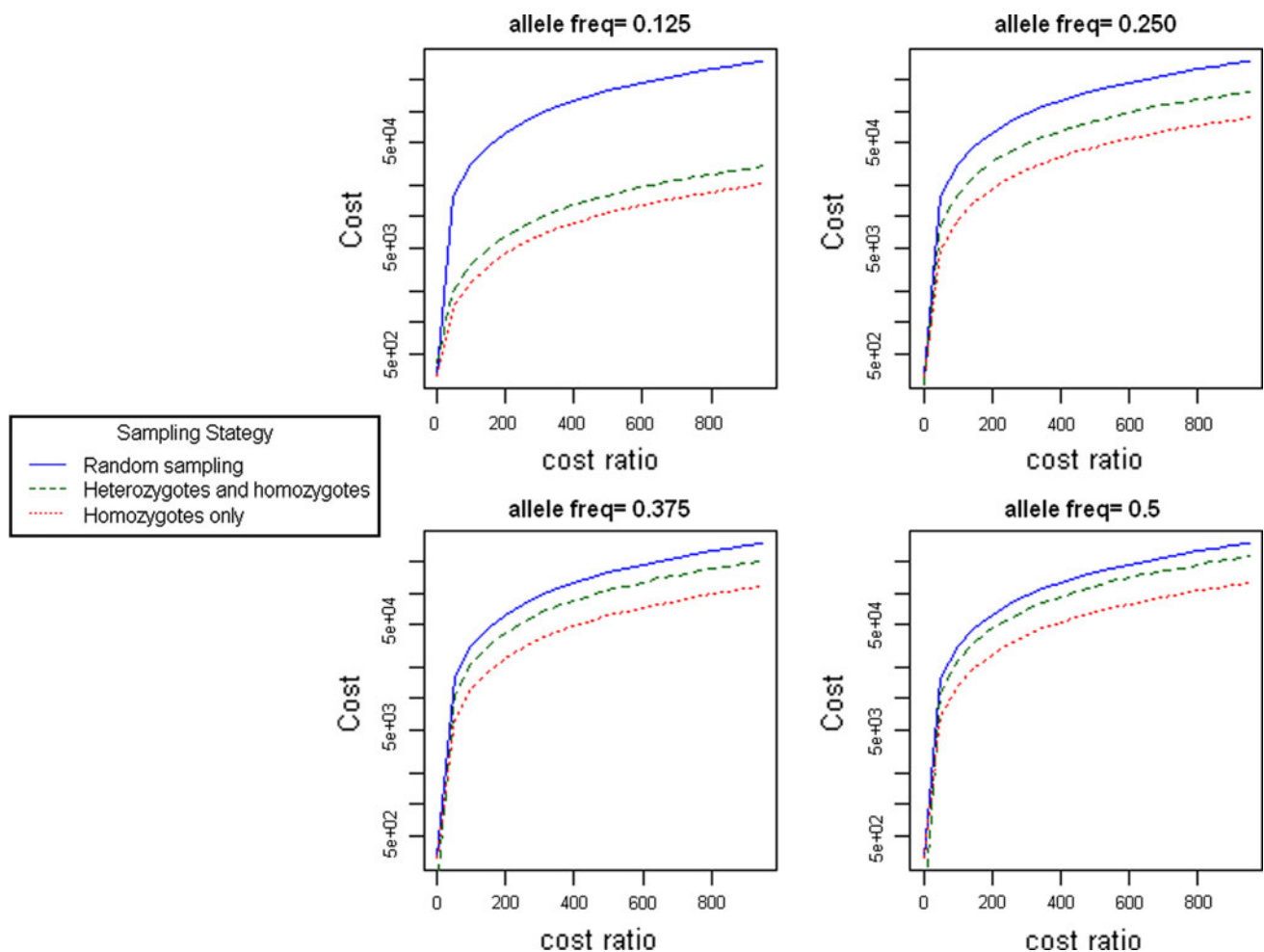
with an explained variance of genotype of 0.03 and a risk allele frequency of 0.3. This effect size is within the range of effect sizes reported by recent genome-wide association studies. As can be seen in Fig. 1, 100 homozygous subjects were sufficient to provide a power of 0.80 for detecting phenotypic differences at a significance level of 0.05. This constitutes a more than fourfold reduction in the sample size required compared to a strategy with random selection. Figure 2 shows the superior power of a sample strategy in which participants were selected for three loci. Illustrative power graphs obtained under different scenarios (explained variances, dominance and allele frequencies) are presented in Figs. 3, 4 and 5. Figure 3 shows that the power is favorable regardless of the amount of explained variance of the genetic variant, but the gain is higher in scenarios where the genetic variant explains only a limited proportion of the phenotypical differences. Figure 4 shows that the gain is highest in scenarios with a fully additive genetic model, while Fig. 5 shows that the gain is highest in scenarios where the allele frequency of the genetic variant is low. This is logical, considering that the impact of selecting for genotype is highest in these situations. The cost-effectiveness is depicted in Figs. 6 and 7 for different ratios of phenotyping and genotyping costs. In the case where the phenotyping cost is \$10 and genotyping is \$1, the ratio is 10. In scenarios where the phenotyping information is already available, the ratio is zero. We here depicted ratios of 0, 200, 700, and 1,000. At present, genome-wide genotyping is less than \$0.001 per genotype and costs



**Fig. 3** The power of sampling strategies under different assumptions of explained variance



**Fig. 4** The power of sample strategies under different assumptions of dominance



**Fig. 5** The power of sample strategies under different assumptions of allele frequencies

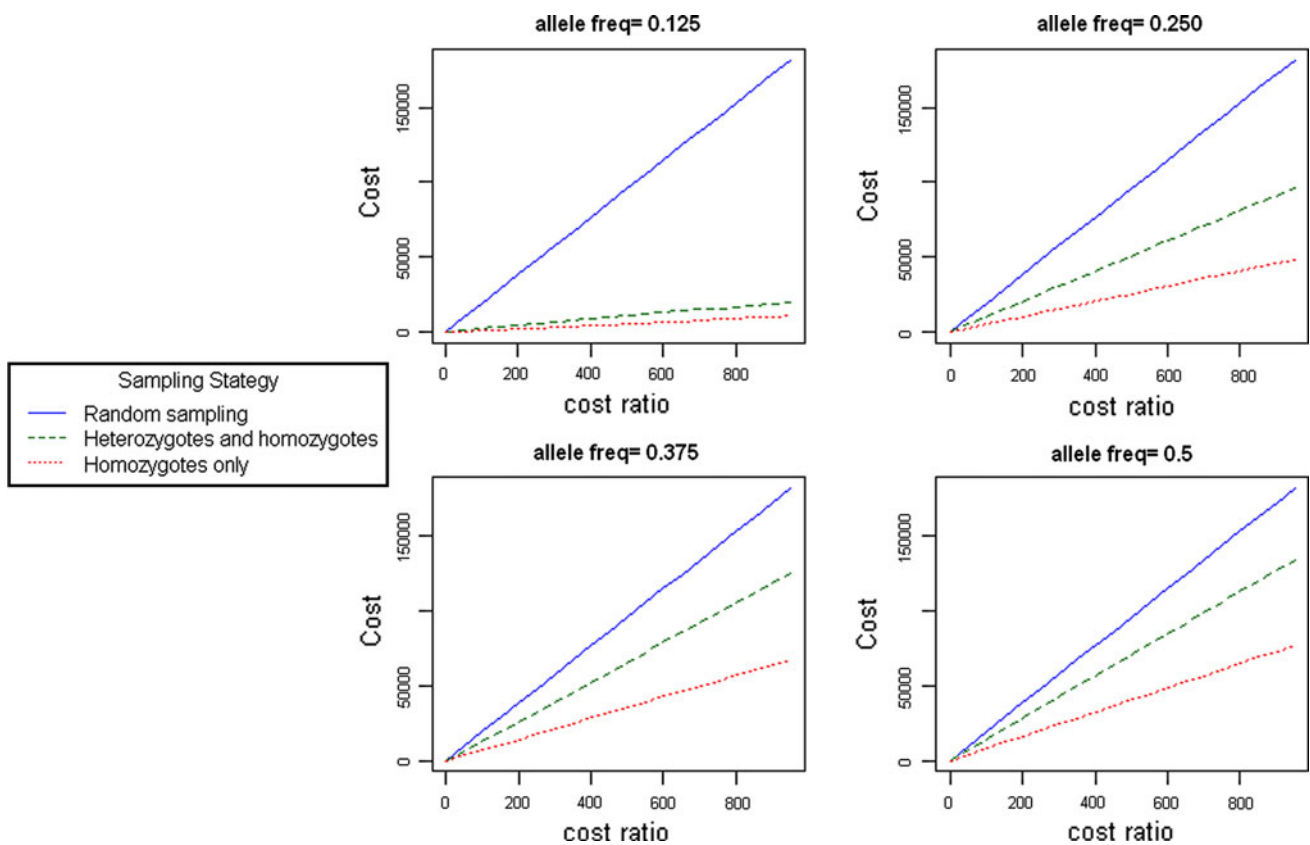
about \$300–\$700 per subject, depending on the platform and reagents; e.g., Taqman for single SNPs costs roughly \$0.20–1. Generally, the true cost of phenotyping is manifold more and should include staff salaries for (clinical) assessments, laboratory work and material expenses. In cases where MRI or PET imaging is involved, phenotype/genotype cost ratios may easily exceed 1,000. Figures 6 and 7 show that the cost reduction is highest in those scenarios in which the allele frequency of the genetic variant is low and increases as the phenotyping costs increase. The cost reduction decreases as the variance explained by the genotype increases (Fig. 7).

## Discussion

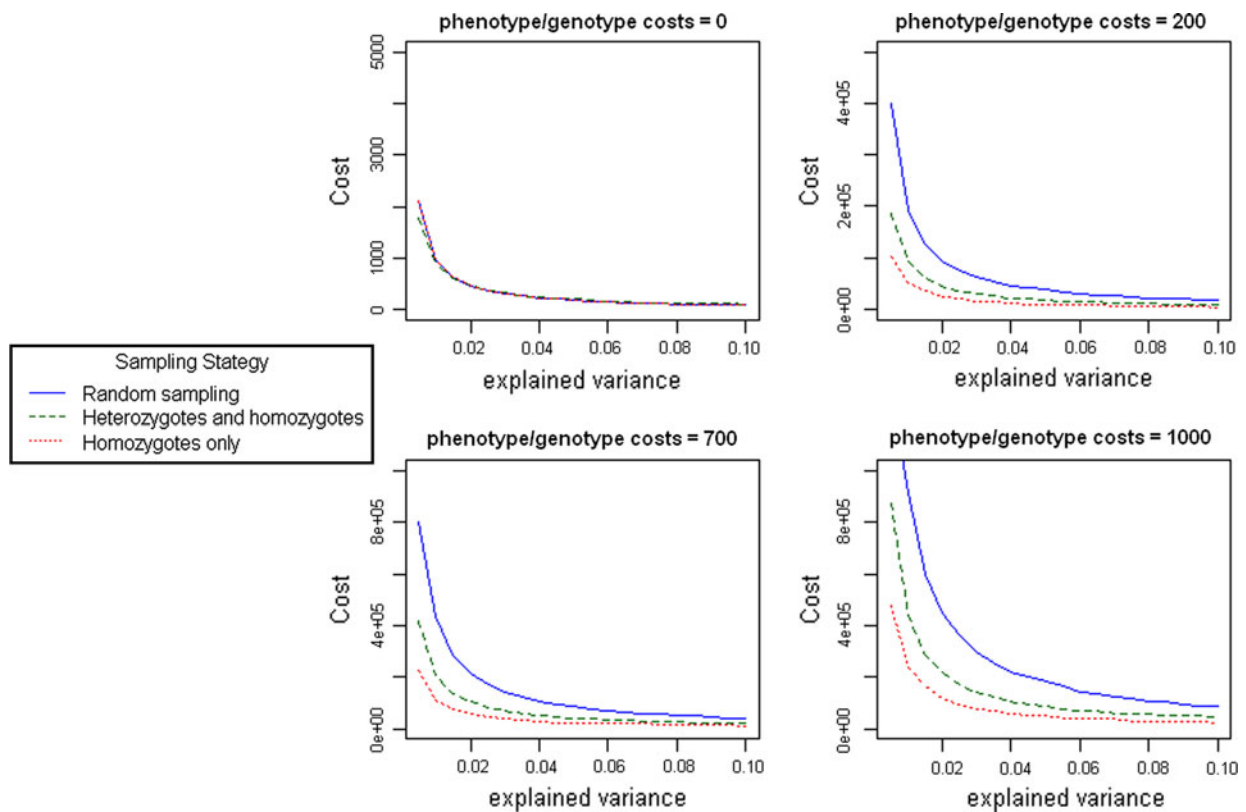
We have demonstrated that “forward genetics” is a cost-effective method for investigating genotype–phenotype

associations in complex human traits. The figures, formulas and R-script provided can be used to estimate the sample size required under different conditions and our calculations show that a marked increase in power and reduction of costs can be obtained by this approach, particularly for those scenarios with relatively high phenotyping costs. Considering the ever reducing costs of genotyping, and the use of increasingly complex phenotypes, such as measures of brain volume obtained with MRI, these scenarios are frequently encountered. If, for instance, after pooling of data on hippocampal volume in a consortium (covering several hundreds of subjects), genome-wide analysis yields a few, significantly associated loci, it would make sense to type these loci in a large population sample and to obtain MRI scans in two groups based on the selection of all these SNPs in order to maximize power. Clearly, such a replication sample could also be used for further genome-wide analysis, but at least the sample would be optimal for replicating the effects found. This method, therefore, is a





**Fig. 6** The costs as a function of phenotype/genotype cost ratio under different allele frequencies



**Fig. 7** The costs as a function of explained variance by the sequence variant under different phenotype/genotype costs ratios

cost-effective way to confirm previously reported genetic association signals, to refine genotype–phenotype relationships, and thus to facilitate the discovery of susceptibility loci in complex quantitative traits. Extensions of this design, in which subjects are selected on the basis of two or more genetic variants, have been presented here. Interaction between the genes can subsequently be tested as a measure of epistasis.

Applying forward genetics can be particularly helpful in studies where DNA is available as part of an epidemiological (longitudinal) study. We plan to use this approach for the study of cognitive and brain volumetric phenotypes in a large, population-based epidemiological sample from the Netherlands. This study will be performed as part of the Utrecht Health Project (Grobbee et al. 2005). In this design DNA of more than 6,000 subjects is available for genotyping. Subjects homozygous for alleles that are associated with cognitive traits, and subjects homozygous for the wildtype allele, will be invited for further comprehensive assessments. Our power analyses show that a number of  $N = 90$  per genetic variant will be sufficient to obtain a type I error rate of 0.007 and a power of 0.8. We plan to comprehensively assess the psychological-, cognitive-, and social functioning of the selected subjects in order to test seven different phenotypes. The raters and participants will remain blind to the subject's genotype status in order to avoid phenotyping bias and ethical complications. The classical approach would be to assess a large numbers of participants from the epidemiological study and to perform a case–control association analysis or quantitative trait locus (QTL) analysis. As pointed out above, the costs of assessing the large numbers of participants required in the conventional design would be much higher compared to the forward genetics approach. Particularly in cases where the genetic variant is rare and many assessments would be required to obtain sufficient power.

Our approach differs from the recently advocated “reverse phenotyping” approach (Schulze and McMahon 2004), in which both genotype and phenotype information is assumed to be available. The phenotype–genotype relationship is, however, re-assessed by revisiting phenotypes (reverse phenotyping). Clearly, a strong argument for Schulze and McMahon's approach (Schulze and McMahon 2004) can be made and there are several recent studies that have successfully applied such reverse phenotyping (Tilley et al. 1998; Silverman et al. 2002; Schaid et al. 2006). However reverse phenotyping cannot solve the power issue, which is relevant to the investigation of rarer genetic variants; nor does it result in a cost reduction in the study of expensive phenotypes, such as those based on brain imaging. Overall, our analyses demonstrate that forward genetics can be a useful tool for studying more complex and expensive

behavioral phenotypes in the context of epidemiological studies.

**Acknowledgements** We are grateful to J.L. Senior for editing the manuscript.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS, The Type 1 Diabetes Genetics Consortium (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet.* 10 May (epub ahead of print)
- Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longmans Green, Harlow, Essex, UK
- Grobbee DE, Hoes AW, Verheij TJ, Schrijvers AJ, van Ameijden EJ, Numans ME (2005) The Utrecht Health Project: optimization of routine healthcare data for research. *Eur J Epidemiol* 20:285–287
- Purcell S, Cherny SS, Sham PC (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19(1):149–150
- Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* 268:1584–1589
- Schaid DJ, McDonnell SK, Zarfes KE, Cunningham JM, Hebring S, Thibodeau SN, Eeles RA, Easton DF, Foulkes WD, Simard J, Giles GG, Hopper JL, Mahle L, Moller P, Badzioch M, Bishop DT, Evans C, Edwards S, Meitz J, Bullock S, Hope Q, Guy M, Hsieh CL, Halpern J, Balise RR, Oakley-Girvan I, Whittemore AS, Xu J, Dimitrov L, Chang BL, Adams TS, Turner AR, Meyers DA, Friedrichsen DM, Deutsch K, Kolb S, Janer M, Hood L, Ostrander EA, Stanford JL, Ewing CM, Gielzak M, Isaacs SD, Walsh PC, Wiley KE, Isaacs WB, Lange EM, Ho LA, Beebe-Dimmer JL, Wood DP, Cooney KA, Seminara D, Ikonen T, Baffoe-Bonnie A, Fredriksson H, Matikainen MP, Tammela TL, Bailey-Wilson J, Schleutker J, Maier C, Herkommer K, Hoegel JJ, Vogel W, Paiss T, Wiklund F, Emanuelsson M, Stenman E, Jonsson BA, Gronberg H, Camp NJ, Farnham J, Cannon-Albright LA, Catalona WJ, Suarez BK, Roehl KA (2006) Pooled genome linkage scan of aggressive prostate cancer: results from the International Consortium for Prostate Cancer Genetics. *Hum Genet* 120:471–485
- Schulze TG, McMahon FJ (2004) Defining the phenotype in human genetic studies: forward genetics and reverse phenotyping. *Hum Hered* 58:131–138
- Silverman EK, Mosley JD, Palmer LJ, Barth M, Senter JM, Brown A, Drazen JM, Kwiatkowski DJ, Chapman HA, Campbell EJ, Province MA, Rao DC, Reilly JJ, Ginns LC, Speizer FE, Weiss ST (2002) Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes. *Hum Mol Genet* 11:623–632
- Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietiläinen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T,

- Børglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Böttcher Y, Olesen J, Breuer R, Möller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Réthelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemenev LA, Genetic Risk and Outcome in Psychosis (GROUP), Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Touloupoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jönsson EG, Terenius L, Agartz I, Petursson H, Nöthen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA (2009) Common variants conferring risk of schizophrenia. *Nature* 460:744–747
- R Development Core Team (2005) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Tilley L, Morgan K, Kalsheker N (1998) Genetic risk factors in Alzheimer's disease. *Mol Pathol* 51:293–304