Genetic Epidemiology 34 : 479–491 (2010)

Design of Association Studies with Pooled or Un-pooled Next-Generation Sequencing Data

Su Yeon Kim,[1]* Yingrui Li,[2] Yiran Guo,[2] Ruiqiang Li,[2] Johan Holmkvist,[3] Torben Hansen,[3,4] Oluf Pedersen,[3,5,6] Jun Wang,[2,7] and Rasmus Nielsen[1,2,7]

[1]*Departments of Integrative Biology and Statistics, UC Berkeley, Berkeley, California*
[2]*Beijing Genomics Institute, Shenzhen, China*
[3]*Hagedorn Research Institute, Gentofte, Denmark*
[4]*Faculty of Health Science, University of Southern Denmark, Odense, Denmark*
[5]*Faculty of Health Science, University of Aarhus, Aarhus, Denmark*
[6]*Institute of Biomedical Science, University of Copenhagen, Denmark*
[7]*Department of Biology, University of Copenhagen, Copenhagen, Denmark*

Most common hereditary diseases in humans are complex and multifactorial. Large-scale genome-wide association studies based on SNP genotyping have only identified a small fraction of the heritable variation of these diseases. One explanation may be that many rare variants (a minor allele frequency, MAF $<5\%$), which are not included in the common genotyping platforms, may contribute substantially to the genetic variation of these diseases. Next-generation sequencing, which would allow the analysis of rare variants, is now becoming so cheap that it provides a viable alternative to SNP genotyping. In this paper, we present cost-effective protocols for using next-generation sequencing in association mapping studies based on pooled and un-pooled samples, and identify optimal designs with respect to total number of individuals, number of individuals per pool, and the sequencing coverage. We perform a small empirical study to evaluate the pooling variance in a realistic setting where pooling is combined with exon-capturing. To test for associations, we develop a likelihood ratio statistic that accounts for the high error rate of next-generation sequencing data. We also perform extensive simulations to determine the power and accuracy of this method. Overall, our findings suggest that with a fixed cost, sequencing many individuals at a more shallow depth with larger pool size achieves higher power than sequencing a small number of individuals in higher depth with smaller pool size, even in the presence of high error rates. Our results provide guidelines for researchers who are developing association mapping studies based on next-generation sequencing. *Genet. Epidemiol.* 34 : 479–491, 2010.

Key words: pooled samples; association mapping; rare allele; optimal design; next-generation sequencing

Contract grant sponsors: The Lundbeck Foundation Centre of Applied Medical Genomics in Personalized Disease Prediction, Prevention and Care (LuCAMP); NIH; Contract grant numbers: R01-HG003229; R01-MH084691.
Additional Supporting Information may be found in the online version of this article.
*Correspondence to: Su Yeon Kim, Departments of Integrative Biology and Statistics, University of California, Berkeley, 3060 Valley Life Sciences Bldg ♯3140, Berkeley CA 94720. E-mail: suyeonkim@berkeley.edu
Received 2 October 2009; Revised 28 January 2010; Accepted 21 February 2010
Published online 15 June 2010 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20501

INTRODUCTION

Determining the genetic basis of complex genetic diseases is one of the main challenges in human genetics [Altshuler et al., 2008; Frazer et al., 2009; Kruglyak, 2008; McCarthy and Hirschhorn, 2008; McCarthy et al., 2008]. Genome-wide association mapping studies (GWAs) have revealed many susceptibility loci for complex diseases such as diabetes, hypertension, bipolar, Crohn's disease and others [Barrett et al., 2008; Hindorff et al., 2009; Thomas et al., 2009; Wellcome Trust Case Control, 2007]. However, despite many successes, a majority of the additive generic variance still remains unexplained [Altshuler et al., 2008; Frazer et al., 2009; Kruglyak, 2008; McCarthy and Hirschhorn, 2008]. One explanation for this result is that rare variants might play a significant role in complex diseases. Since genotyping platforms include SNPs that are discovered by sequencing a small panel, most of the included SNPs are common variants [MAF $>5\%$; Frazer et al., 2007] in previous GWAs. The importance of rare variants in understanding complex traits has been discussed in a number of recent studies [Azzopardi et al., 2008; Bodmer and Bonilla, 2008; Cohen et al., 2004, 2006; Gorlov et al., 2008; Iyengar and Elston, 2007; Ji et al., 2008; McClellan et al., 2007; Polychronakos, 2008]. For example, in a re-sequencing study of three genes, Ji et al. [2008] found several rare mutations related to hypertension, appearing in the heterozygous state.

The cost of re-sequencing has dropped dramatically over the past few years [Bentley et al., 2008; Wang et al., 2008; Wheeler et al., 2008]. This recent development has made next-generation re-sequencing a viable alternative to SNP genotyping based on chips or other platforms

[Bentley et al., 2008; Hillier et al., 2008; Van Tassell et al., 2008]. However, in most studies full genome-wide re-sequencing at a high depth is still prohibitively expensive. Instead, two strategies are being used to reduce costs: (1) sequencing of the exome only—using various exon-capturing techniques and (2) the use of pooled samples. With recent developments in exon-capturing techniques, most of the human exons can be efficiently captured [Albert et al., 2007; Gnirke et al., 2009; Hodges et al., 2007; Krishnakumar et al., 2008; Okou et al., 2007; Porreca et al., 2007]. While many functional variant may lie in non-coding regions, many of the mutations with the best potential for translational medicine are in coding regions, providing an extra motivation for focusing on the exome [Jones et al., 2009; Kryukov et al., 2009; Raymond et al., 2009].

The use of pooled samples is particularly attractive in two-stage designs [Chi et al., 2009; Skol et al., 2006; Wang et al., 2006; Zuo et al., 2006, 2008]. In the first stage of a two-stage design, a large number of markers are genotyped, or re-sequencing is being used, on a moderate number of individuals. In the second stage, the most promising markers from the first stage are further examined by genotyping in a larger set of individuals. To reduce the cost of large-scale association studies, pools of DNA from many individuals have been successfully used in the first stage of the two-stage design [Bansal et al., 2002; Boss et al., 2009; Nejentsev et al., 2009; Norton et al., 2004; Sham et al., 2002]. The use of pooled samples reduces cost, particularly when combined with techniques such as exon-capturing, which has a cost associated with each pool to which it is applied. With the advent of cheap re-sequencing techniques, a viable, an economically attractive protocol for GWAs may include an initial stage of exon-capturing and re-sequencing in pooled samples [Druley et al., 2009; Ingman and Gyllensten, 2008; Jones et al., 2009; Lavebratt and Sengul, 2006], and a second stage based on genotyping of the best candidate SNPs from the first stage. Such a protocol is cost effective and has the potential to detect rare SNPs that would not be captured by any of the major genotyping platforms.

Here, we examine the feasibility of this protocol, using both pooled and un-pooled samples. There are a number of issues related to this protocol that warrants further research, including the effects of sequencing errors and pooling variance, and the consequences of these factors for the choice of optimal experimental protocol.

We first make a limited experimental study based on Nimblegen®exons capturing, and sequencing using the Illumina genome analyzer®, to determine the feasibility of studies based on a combination of these techniques. We then use the results of this study to model the protocol using realistic parameters. We perform simulations to determine power and to identify optimal designs.

METHODS

EXPERIMENTAL DESIGN

Our analyses assume a two-stage design, where the first stage may include exon-capturing and/or pooled samples, and will involve next-generation re-sequencing. The second stage is based on traditional SNP genotyping of the most promising variants from the first stage. While the properties and design of the second stage have been explored extensively in the previous work [Skol et al., 2006; Zuo et al., 2008], the use of next-generation sequencing for the first stage [e.g., Nejentsev et al., 2009] has not previously been explored to the same degree.

EXPERIMENTAL DATA

A small empirical study was conducted to evaluate feasibility of this novel approach and assess the variability introduced in real experiments. Genomic DNA was purified from blood leucocytes from five Danish volunteers recruited at Steno Diabetes Center, Denmark. The volunteers gave informed consent and the research protocol, which is focused on studies of the genetics of metabolic disorders, was approved by the local Ethical Committee of Copenhagen. Two pools were constructed, one with two individuals and the other with five individuals. Pooling DNA of individuals was done following the pre-PCR protocol described in Lavebratt and Sengul [2006] with some small modification. In our procedure, individual DNA was diluted to 10 ng/μl instead to 5 ng/μl before pooling. Our pooling procedure is summarized in Supplementary Figure 1. Exon-capturing was performed using a NimbleGen chip® [Albert et al., 2007], which captures 6,726 "exonic" regions covering ~5 Mb sequence. The DNA sample in each pool was fragmented to an average size of 500 bp, end repaired, ligated to Solexa adaptors®, hybridized and captured using the high-density oligonucleotide microarray. Finally, the DNA sample was amplified using PCR, and the eluted sample from it was sequenced using Solexa Genome Analyzer II®. Solexa sequence reads were aligned to the reference human genome (NCBI build 36) using read-alignment program SOAP [Li et al., 2008].

Pooling efficacy. In our design, reads from individuals in pools could be identified after re-sequencing because they were re-sequenced both individually and in pools. The individual re-sequencing allowed the identification of unique mutations specific to each individual. Pooling efficacy was then measured by the degree to which the pooled DNA was composed of an equal contribution of DNA from each individual.

Consider a pool with P_s number of individuals. When using pooled samples, individual DNA samples are measured and pooled, then a fixed amount of pooled DNA is prepared for sequencing (Supplementary Fig. 1). Therefore, the sequencing depth of each individual is not controlled directly, but is a function of the proportion of DNA from each individual and the total sequencing depth. However, the mean depth for an individual is expected to be proportional to the amount of DNA of that individual in the pool. Therefore, one way to model the pooling efficacy is as follows:

$$Y_{i,B}, Y_{i,b} \sim \text{Poisson}(\mu \times R_i), \text{ where } \mu = \frac{D_p}{2} \text{ and}$$
$$R_i = \frac{W_i}{\sum_j W_j}, W_i \sim \Gamma\left(\alpha_p, \frac{1}{\alpha_p}\right), \quad (1)$$

where $Y_{i,B}$ and $Y_{i,b}$ are the counts of read bases generated from each of the two alleles (B and b) in individual i, D_p is the pool depth and R_i is the proportion of the DNA from each individual. W_i is the DNA amount of individual i. Note that if pooling is perfect, then W_i is the same across individuals and R_i is expected to be $1/P_s$. In reality, however, W_i is likely to vary, and one way to model this

variation is to use a gamma distribution, as the gamma distribution provides a very flexible family of continuous positive variables using only two parameters. The gamma distribution is a favorite choice in statistics for modeling such distributions and the properties of the distribution are well known. In our case, when W_i follows a gamma distribution with a shared rate parameter, $(R_1, \ldots, R_{P_s}) \sim \text{Dirichlet}(\alpha_p)$. Note that the mean and the variance of each component in the Dirichlet distribution is:

$$E(R_i) = \frac{1}{P_s} \quad \text{and} \quad \text{Var}(R_i) = \frac{P_s - 1}{P_s^2 \times (\alpha_p P_s + 1)} \qquad (2)$$

The parameter α_p controlling the pooling variance is estimated from the data using the maximum likelihood method.

Exon-capturing efficacy. Variation in exon-capturing efficacy introduces extra variation in read depth across the sites. High depth implies efficient exon-capturing, and low depth implies poor exon-capturing performance. We model the exon-capturing efficacy as a gamma distribution:

$$Y_j \sim \text{Poisson}(\mu_t \times C_j) \quad \text{and} \quad C_j \sim \Gamma\left(\alpha_c, \frac{1}{\alpha_c}\right), \qquad (3)$$

where Y_j is the observed total depth at site j and μ_t is the expected total sequence depth given by experimental design. C_j controls the exon-capturing efficacy at site j. Large values of C_j imply efficient capturing on site j and small values C_j imply poor exon-capturing efficacy. Exon-capturing might fail for some sites. In that case, the capturing efficacy can be modeled as a mixture of a point mass at zero and a gamma distribution.

We estimate the sequencing error rate conservatively as the average mismatch rate. Specifically, all the counts of pairwise mismatches between the reference human genome and aligned reads were averaged across targeted regions as well as across the aligned reads.

EXPERIMENTAL COST

In order to illustrate how to make cost effective choices regarding optimal design, we will make assumptions regarding experimental cost. Obviously, costs can change rapidly and design of individual studies should take current costs into consideration. It should be noticed that our general conclusions are not dependent on the specific assumptions regarding costs made here. In the most general pooled design, the total cost of an experiment has four components: cost of obtaining DNA samples, cost of pooling DNA samples, exon-capturing, and sequencing. We will in the following ignore the cost of obtaining DNA samples and assume that such samples are available before the onset of the study. The DNA pooling cost mainly depends on the total number of individuals that are pooled. The key step in pooling is to dilute and accurately quantify the DNA concentration of each individual sample. We will here assume, for the sake of example, that initial preparation for each individual in a pool costs \$2. We will assume that exon-capturing covering all human exons (~30 Mb) costs \$3,500 for each pool. The cost of sequencing is divided into two parts, a cost of preparation of the DNA sample which we will assume is \$200, and a cost associated with the production of sequencing reads. We will assume a cost of \$500 for Solexa-sequencing of a 30 Mb exonic region at a depth of 2× (for details, see Supplementary Methods). The assumed experimental costs are summarized in Table I. Notice that for individual sequencing, the exon-capturing cost rapidly increases with the number of individuals, even if the sequencing cost is fixed. The use of pooled samples allows deeper sequencing at the same cost.

LIKELIHOOD RATIO TEST

The data produced by next-generation sequencing differ from that of SNP-chips. Genotyping, in principle, reveals the two alleles in each individual in each targeted nucleotide site. Next-generation re-sequencing produces large amounts of short reads. After mapping to the reference genome, an alignment of reads across the targeted regions is obtained. A schematic example of re-sequencing data at a single site is shown in Figure 1. In this particular example, cases and controls each consist of two pools with two individuals. Each nucleotide site in each individual is represented in reads a random number of times. It is generally unknown which of the two alleles in an individual is represented in a particular read. Also, there is a high probability of sequencing errors.

One of the main challenges in the use of next-generation sequencing data in association mapping studies aimed at detecting associations with rare SNPs is to distinguish between true SNPs and false SNPs caused by sequencing

TABLE I. Examples of experimental costs for re-sequencing of pooled or un-pooled DNA samples in targeted regions

Number of cases/controls	Pool size	Depth per indiv.	Experimental cost (unit: \$1,000)			
			Pooling	Exon-capturing	Sequencing	Total
500	1	8×	0	3,500	2,200	5,700
1000	1	4×	0	7,000	2,400	9,400
2000	1	2×	0	14,000	2,800	16,800
500	1	8×	0	3,500	2,200	5,702
1000	2	4×	4	3,500	2,200	5,704
2000	4	2×	8	3,500	2,200	5,708
4000	8	1×	16	3,500	2,200	5,716
1000	5	4×	4	1,400	2,080	3,484
1000	5	8×	4	1,400	4,080	5,484
1000	5	16×	4	1,400	8,080	9,484

The cost settings are obtained by assuming Solexa-sequencing of pooled or unpooled DNA samples on the human exome (~30 Mb), captured using Nimblegene arrays.

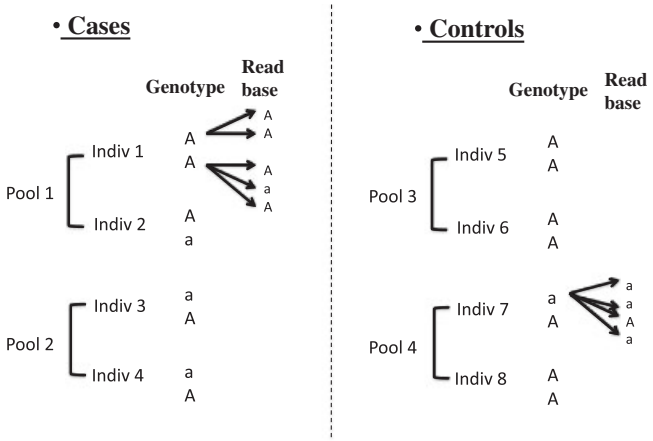

Fig. 1. Schematic illustration of the next-generation sequencing data at a single position. Each case and control sample consist of two pools with two individuals in each pool. The two alleles of each individual are shown in the "Genotype" column. There are two types of alleles (*A* and *a*). Each allele appears a random number of times, and may have been affected by sequencing errors.

errors. The nature of the data is such that, even after basic bioinformatic quality control procedures, sequencing errors remain a serious problem. Power can be gained by incorporating the possibility of errors into the statistical framework used for detecting associations. There is a trade-off between eliminating too many putative SNPs as sequencing errors, thereby eliminating a number of true SNPs, and eliminating too few putative SNPs, and then suffer a serious multiple testing problem. A full likelihood approach incorporated into the association mapping testing procedure may help to directly identify the optimal balance between eliminating too few or too many putative SNPs in a particular design, and to select the putative SNPs that most likely are associated with the trait. We have developed a simple version of such a likelihood approach. We have also extended this method so it can be applied to provide a more powerful method for pooling designs.

In the following we will discuss how the method can be developed into a likelihood ratio test, which we apply for each individual site to test the difference in minor allele frequencies: $H_0 : p_1 = p_2 (= p_0)$ vs. $H_A : p_1 \neq p_2$, where p_1 and p_2 are the minor allele frequency in cases and controls, respectively.

We test each site, j, independently of each other. Suppose that for both cases and controls, there are N_{pool} pools with P_s individuals per pool. After re-sequencing, for each pool m, an alignment of reads $O^{(m)} = (X_1^{(m)}, \ldots, X_{V^{(m)}}^{(m)})'$ is obtained, where $X_k^{(m)}$ is the kth read among the total $V^{(m)}$ reads in locus j. Let $G^{(m)}$ be the number of A (minor) alleles in pool m. (Note that $G^{(m)}$ is not observed.) Each observed read is a copy of one of the alleles in a pool, but copying is potentially made with errors. Note that below, we assume a relatively simple structure in which $\varepsilon = P(\text{read} = A|\text{allele} = a) = P(\text{read} = a|\text{allele} = A)$. However, our statistical model can easily be extended to incorporate a more complicated error structure. In Equation (4) below, instead of a binomial distribution modeling the conditional read counts a multinomial distribution with probability that takes each type-specific error rate ($\varepsilon_{b,b'}$) into account could be used: $\varepsilon_{b,b'} = P(\text{read} = b|\text{allele} = b')$, where $b, b' \in (A, C, G, T)$.

Our likelihood ratio statistic (*LRT*) is computed as:

$$\begin{aligned} LRT &= -2\log\left(\frac{L(\hat{p}_0|O,\varepsilon)}{L(\hat{p}_1,\hat{p}_2|O,\varepsilon)}\right) \\ &= -2\log\left(\frac{\prod_{m=1}^{2N_{\text{pool}}} P(O^{(m)}|\hat{p}_0,\varepsilon)}{\prod_{m=1}^{N_{\text{pool}}} P(O^{(m)}|\hat{p}_1,\varepsilon)\prod_{m=N_{\text{pool}}+1}^{2N_{\text{pool}}} P(O^{(m)}|\hat{p}_2,\varepsilon)}\right), \end{aligned}$$

and, the likelihood for each pool m is computed as:

$$\begin{aligned} P(O^{(m)}|p_0,\varepsilon) &= \sum_{G^{(m)}} P(G^{(m)}|p_0)\, P(O^{(m)}|G^{(m)},\varepsilon) \\ &= \sum_{k=0}^{S_{\text{pool}}} \left(P(G^{(m)} = k|p_0) \prod_{r=1}^{V^{(m)}} P(X_r^{(m)}|G^{(m)} = k,\varepsilon)\right) \\ &= \sum_{k=0}^{S_{\text{pool}}} \Bigg(\text{Binom}(k; S_{\text{pool}}, p_0)\text{Binom}(n_A^{(m)}; V^{(m)}, \\ &\qquad \frac{k}{S_{\text{pool}}}(1-\varepsilon) + \left(1 - \frac{k}{S_{\text{pool}}}\right)\varepsilon\Bigg), \end{aligned} \tag{4}$$

where $n_A^{(m)}$ is the number of A in reads. The last equation is motivated as follows: $G^{(m)}$ follows a binomial distribution with S_{pool} ($= 2P_s$) number of trials each with probability p_0 (assuming no population structure). Given that there are k minor alleles (A) in a pool, each read becomes of type A with probability k/S_{pool} if there are no errors. If errors occur, the probability is

$$\frac{k}{S_{\text{pool}}}(1-\varepsilon) + \left(1 - \frac{k}{S_{\text{pool}}}\right)\varepsilon$$

Assuming homogeneity in sequencing efficiency across chromosomes (i.e., the number of sequenced reads for each allele follows the same distribution) and independence, the number of A reads in the pool ($n_A^{(m)}$) follows a binomial distribution. Notice that the total number of reads in each pool ($V^{(m)}$) is assumed to be given.

The performance of the likelihood ratio statistic for hypothesis testing based on individual sites could be evaluated in terms of Type I and Type II errors in a classical sense. However, our objective is to evaluate design issues relating to a two-step design in which next-generation sequencing is used in the first step of the design. We therefore evaluate the efficacy of the statistic in selecting a causal SNP among a pre-determined number of selected sites. Accordingly, the power is determined as the probability of including the causative SNP among the set of selected sites (e.g., the probability of including the causal SNP among 50 sites selected from a total of 300,000 sites). We do not evaluate the Type I and Type II error of the test in the classical sense, although we will present simulation results showing that *P*-values calculated based on the likelihood ratio test statistic under suitable conditions follow a uniform distribution, and hence, that the test will have a Type I error rate equal to the nominal significance level.

Maximization of the likelihood function was done using a bounded version of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970] or the Simplex method [Nelder and Mead, 1965]. Overall, the Simplex method algorithm performed better, presumably because the computational expense in numerically approximating the gradient in the BFGS algorithm overwhelmed the faster convergence of this algorithm for this low-dimensional problem. Computing our LRT statistic for 1,000 sites takes 69 sec with the Simplex algorithm, but 136 sec with BFGS algorithm on a standard desktop computer, for a total of 10,000 reads from 500 cases and 500 controls in 1,000 pools.

G-TEST

With genotyping data and under the Hardy-Weinberg assumption, one natural way to test for differences in allele frequencies between cases and controls is to use *G*-tests. *G*-tests are likelihood-ratio tests, which are commonly used to test independence in contingency tables. When counts are generated by independent trials, the test statistic asymptotically follows a χ^2 distribution. The test statistic, *G*-statistic, is computed as:

$$G = 2\sum_{i,j} O_{i,j} \log\left(\frac{O_{i,j}}{E_{i,j}}\right) \tag{5}$$

where $O_{i,j}$ is the frequency observed in a cell, and $E_{i,j}$ is the frequency expected under the null hypothesis in which rows and columns are independent in a contingency table. The well-known Pearson's χ^2 test is asymptotically equivalent to *G*-test.

With next-generation sequencing data, the *G*-statistic can be calculated based on the read counts. However, note that then *G*-statistic does not take pooling structure as well as sequencing errors into account. Also, the read counts are not generated independently, since a single allele can be copied multiple times as read bases. As such, the likelihood function on which the *G*-statistic is based is misspecified. In our study, we examined *G*-test mainly to compare the performance with our LRT statistic.

SIMULATING DATA

Assuming a single causative SNP, we performed extensive simulations to examine the statistical properties of the method. In these simulations, the power was evaluated as the probability of including the causative SNP among the set of SNPs selected for the second stage in a two-stage design. We then varied (1) the sequencing error rate per base pair, (2) the number of cases and control individuals, (3) the pool size (the number of individuals per pool), and (4) the sequencing depth per individual. To simplify the simulation procedure, we do not attempt to take population structure and linkage disequilibrium across loci into account. The general conclusions regarding relative power of different designs should not be affected by the strength of LD. For simplicity, we use equal numbers of cases and controls for all of our simulations, so, in the subsequent description, the number of total individuals is twice the number of cases (if not otherwise stated).

Each data set consists of aligned reads from a 300 Kb region, which contains one causative SNP. The null set consists of two subsets, a set of true SNPs and a set of "false" SNPs. False SNPs are sites that are invariable in the sample but appear polymorphic due to sequencing errors. Based on the results of the empirical study, we will assume an error rate of 1%, if not otherwise stated. We compute the number of true SNPs in the region assuming a mutation rate of 5×10^{-4}, as estimated in coding regions by Wang et al. [2008]. For example, with 1,000 cases and 1,000 controls, the number of SNPs is computed as $5 \times 10^{-4} \times \sum_{i=1}^{2,000-1} \frac{1}{i} \times 300,000 = 1,227$. For each true SNP under the null, the minor allele frequency (MAF) is drawn from the distribution of sample frequencies under the assumption of a standard coalescence model. Genotypes are simulated assuming Hardy-Weinberg equilibrium and the reads are generated by copying each allele a Poisson distributed number of times with mean equal to half the per-individual depth. False SNPs are simulated by assuming a MAF of zero. Since every site has high sequencing depth and the next-generation sequencing error is relatively high, approx. 1%, almost all the invariable sites appear as false SNPs in the read alignment. The causative SNP is simulated similarly, but with different MAFs for cases and controls computed using a multiplicative disease model.

When necessary, we also simulate data with variation in pooling efficacy and exon-capturing efficacy. When both effects are introduced, we use the following model:

$$Y_{i,j,b} \sim \text{Poisson}(\mu \times R_i \times C_j) \tag{6}$$

$$\mu = \frac{D_p}{2}, \quad R_i = \frac{W_i}{\sum_i W_i}, \quad W_i \sim \Gamma(\alpha_p, \frac{1}{\alpha_p}), \quad C_j \sim \Gamma(\alpha_c, \frac{1}{\alpha_c}),$$

where $Y_{i,j,b}$ is the count of the reads generated from an allele in individual i at locus j. D_p is the depth of each pool, and R_i is the proportion of the amount of DNA from individual i within a pool. W_i models the amount of DNA from individual i and C_j models exon-capturing efficacy at site j. α_p and α_c controls the pooling variance and capturing variance, respectively, the pooling variance equals $1/\alpha_p$, and the capturing variance equals $1/\alpha_c$. Notice that Equations (1) and (3) are simplified versions of this model, in which only one of the two effects is introduced. When the exon-capturing efficacy is constant among sites, the model simplifies to Equation (1). When the exon-capturing efficacy varies among sites, the expected read count of each allele, and thus the total read count, change in proportion to C_j.

RESULTS

EXPERIMENTAL DATA

We examined seven DNA samples, consisting of data from five Danish individuals, sequenced individually, and in two pools: a pool with two individuals and another pool with five individuals. The individual sequencing data were used to detect unique mutations, which allow us to identify individuals of the corresponding reads. Exon-capturing was performed for each of the seven DNA samples, using a NimbleGen chip, and was subsequently sequenced using Solexa-sequencing (for details, see Methods). We evaluated (a) the performance of DNA pooling, (b) exon-capturing efficacy, and (c) sequencing error rate.

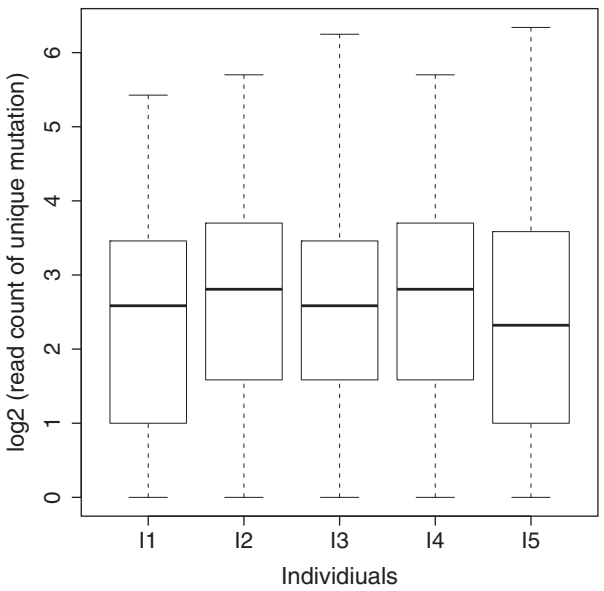

Fig. 2. Pooling variance observed in the empirical study in the pool of five individuals. Using individual sequencing data, unique mutations for each of the five individuals were obtained, and the number of reads with each mutation was counted in the pooled sample. The box plots shows the lower quartile, median (black line), and upper quartile of the log2 of the number of reads with each mutations (*y*-axis) for each individual (*x*-axis). The variation among individuals is relatively small compared to the variation within individuals.

Both pools show good pooling efficacy, i.e., an approximately equal amount of DNA from each individual in the pool (Fig. 2). We first identified mutations that are specific to each individual. Using such mutations, we then estimated the relative amount of DNA from each individual by averaging the number of reads from the unique mutations across the sequenced region.

In the pool with five individuals, 28,026 reads can be uniquely assigned to an individual based on 3,217 diagnostic mutations. The estimated proportion of DNA from each individuals was 0.186, 0.210, 0.191, 0.211, and 0.202, respectively, showing that the proportions differ only slightly. Assuming the model in Equation (1), the estimated pooling variance ($1/\alpha_p$) is $\sim 1/300$, which is very small. The pool with two individuals shows even better performance (data not shown).

The exon-capturing efficacy varies significantly across the genome compared to the Poisson distribution expected under a constant capturing efficacy across the genome (Supplementary Fig. 2). The estimated exon-capturing variance is approx. 1/2 (i.e., α_c is 2) (see Equation 3). Nonetheless, exon-capturing efficacy across samples were homogeneous (Supplementary Fig. 3).

On average, a sequencing error rate of 0.9% was estimated in the Illumina/Solexa raw reads (Table II) using the method described in Methods.

DISTRIBUTION OF THE LRT

We examined the distribution of the likelihood ratio statistic using simulations.

As suggested by standard asymptotic theory, the distribution of the test statistic closely follows a χ^2 distribution with one degree of freedom, except in cases in which the parameters are close to the boundary (Supplementary Fig. 4). Across a range of MAFs covering 0.05–0.5%, the distribution of *p*-values computed based on the χ^2 distribution is nearly uniform. However, when the MAF is very small relative to the assumed sequencing error rate, a sharp peak near 1.0 appears in the distribution of *P*-values, implying an excess of test statistics with nearly zero values. This is due to difficulty in distinguishing true SNPs with very low minor alleles from false ones, and thus MAFs under both the null and the alternative hypothesis are estimated to be zero.

We also varied the number of individuals in a pool (pool size) as well as the sequencing depth per individual. Again, the distribution of the LRT statistic closely follows the χ^2 distribution across a range of pool sizes and

TABLE II. Summary of the empirical data consisting of sequences from seven individuals in 5 MB of exonic DNA

	Indiv. 1	Indiv. 2	Indiv. 3	Indiv. 4	Indiv. 5	Pool 1 1+2	Pool 2 1+2+3+4+5
Sequencing error rate (%)	0.93	0.89	0.89	0.86	0.96	1.05	0.86
Generated reads (Mb)	92.9	86.9	111.5	112.1	104.2	110.0	122.9
% of reads mapped to exons	47.4	44.6	47.6	47.0	52.6	49.8	49.6
Average coverage for the targeted region	18.01	16.86	21.6	21.73	20.22	21.27	21.24

The seven samples consist of five individual samples (Indiv 1-Indiv 5) and two pooled samples, one with a pool size of two (Pool 1) and the other with a pool size of five (Pool 2). Sequencing error rates are conservatively estimated as the average mismatch rate. ''Generated reads'' is the number of the reads after filtering out contaminated reads. ''The percentage of reads mapped to exons'' is the fraction of the reads mapped to exons among the generated reads. ''Average coverage for the targeted region'' is the average number of reads per base-pair across targeted exon regions.

sequencing depths, except for the boundary cases (Supplementary Figs. 5 and 6). When the MAF is very small, the distribution is affected by the pool size and the sequence depth. With increased pool size but constant total read depth, the peak in the distribution becomes more pronounced (Supplementary Fig. 5), because it becomes more difficult to distinguish false SNPs from true SNPs. As sequencing depth increases, the peaks tends to be less pronounced, since high depth helps especially when only a few minor alleles exist (Supplementary Fig. 6).

In summary, the LRT statistic has good statistical properties, in the range of MAF larger than 0.5%, can be used for testing for association, and will form the basis for our analyses of designs. Note that we aim to detect rare alleles with MAF larger than 0.5%. For SNPs with a very small MAF, for example, 0.1%, it would be infeasible (very low power) to detect the causative SNP using 1,000 cases and 1,000 controls.

POWER OF LRT STATISTIC

To examine the power of different experimental designs, we compared the power (computed based on our LRT statistic) varying several parameters: the number of individuals in cases n_1, pool size P_s, individual depth D_i and sequencing error rate ε, pooling variance $1/\alpha_p$, and the capturing variance $1/\alpha_c$ (for detailed definition, see Methods). Note that for computational reasons, we computed the power by selecting 50 SNPs with the strongest association out of 300,000 sequenced sites. This corresponds to selecting 5,000 candidate SNPs out of the human exome (approx. ~30 Mb) for further genotyping. The power is computed based on 5,000 repeated simulations.

For simplicity, we will denote a disease variant with minor allele frequency (MAF) of m and relative risk (RR) of r as $D_{m,r}$.

Effect of sample size. As expected, increasing sample size increases power significantly, across a range of MAFs and relative risks (Fig. 3A). For example, sequencing 500 cases at 4 × with a pool size of five individuals, the power to detect a $D_{1\%,2}$ is 29.8%. However, increasing the sample size by two-fold (1,000 cases) and four-fold (2,000 cases), the power becomes 53.8 and 84.3%, respectively. Also, the power to detect a $D_{5\%,1.5}$ is 43.2, 72.8, and 95.7%, using 500, 1,000, and 2,000 cases.

Effect of pool size. By pooling DNA samples, the cost of exon-capturing can be reduced. However, reads from pools cannot in general be assigned to individuals without additional labeling, potentially leading to a reduction in statistical power and accuracy. To explore this issue, we performed simulations for different pool sizes, but with a fixed number of individuals and a fixed sequencing depth. As expected, the power decreases as pool size increases (Fig. 3B). Nonetheless, the loss in power with increased pool size is relatively small. For example, the power to detect a $D_{1\%,2}$ is 64.8%, with sequencing 1,000 cases individually at 4 ×. By pooling five individuals or even 100 individuals, the power is only reduced to 53.8 and 45.5%, respectively. These results assume that pools with an equal amount of DNA from each individual can be constructed. However, even when using a pooling variance that is 100-fold of the estimated variance from the experimental data, the results remain almost the same (Supplementary Fig. 7).

Effect of sequencing depth. The sequencing depth is the average number of reads covering a site. As expected, the power increases with increasing depth (Fig. 3C). For example, the power to detect a $D_{1\%,2}$ is 33.0%, using sequencing of 1,000 cases at 2 × with a pool size of five. However, with a depth of 4 ×, the power is 53.8%, and with depth of 8 ×, the power is 72.9%. The increased power at higher sequencing depth is caused by an increased ability to distinguish between true SNPs and sequencing errors. The advantage of a higher sequencing depth will, therefore, also depend on the error rate.

Effect of sequencing error rate. The sequencing error rate is one of the major concerns in association studies targeting rare mutations based on re-sequencing. Nonetheless, we find that the sequencing error rate has a relatively small effect on the power, when the error rate is similar or lower than the MAF of the disease SNP (Fig. 3D). For example, even if the error rate decreases by half, from 1 to 0.5%, the power to detect a disease variant with MAF of 1% or more increases only slightly, at most 5%. This also suggests that our LRT statistic properly takes the sequencing error rate into account. With an error rate of 2%, there is a more pronounced decrease in power, especially for a MAF of 1% or less (Fig. 3D). However, in most studies, the error rate will probably be at 1% or below.

Effect of pooling performance. Our simulation results suggest that the effect of the pooling variance is of minor concern, for the pool sizes considered here. The pooling variance estimated from our empirical study corresponds to $V_p = 1/300$, and the power with this pooling variance is nearly the same as with a variance equal to zero (Fig. 3E). Even if the variance increases by 100-fold, the power decreases only slightly (Fig. 3E). For example, the power to detect a $D_{1\%,2}$ is 53.8% without pooling variance, and it is 51.8% with a pooling variance of 1/3.

Effect of exon-capturing efficacy. Exon-capturing efficacy may vary across sites. In our empirical study, the total depth varied significantly across the sites, and the estimated exon-capturing variance was 1/2. This leads to a markedly reduced power compared to the case of almost no variation in exon-capturing efficacy ($V_c = 1/100$) (Fig. 3F). For example, the power to detect a $D_{1\%,2}$ is 53.9% without exon-capturing variance, while the power is 24% with $V_c = 1/2$. The major reason is that with a high exon-capturing variance, there is an increased probability of low sequencing depth in the site containing the causal SNP.

POWER OF THE LRT WITH ESTIMATED ERROR RATE

Our LRT statistic is computed assuming a known sequencing error rate. In practice, the sequencing error rate will be estimated from the data, possibly by averaging across sites in a region assumed to have a constant error rate. Mis-specification of the error rate may reduce the performance of our LRT statistic. Supplementary Figure 8 shows that mis-specification of the error rate as 0.5% when the true error rate is 1% decreases the power significantly. However, the loss of power due to mis-specification of the error rate is less of our concern, since the error rate can be estimated at each site under the null hypothesis. With thousands of reads, an error rate close to 1% can be

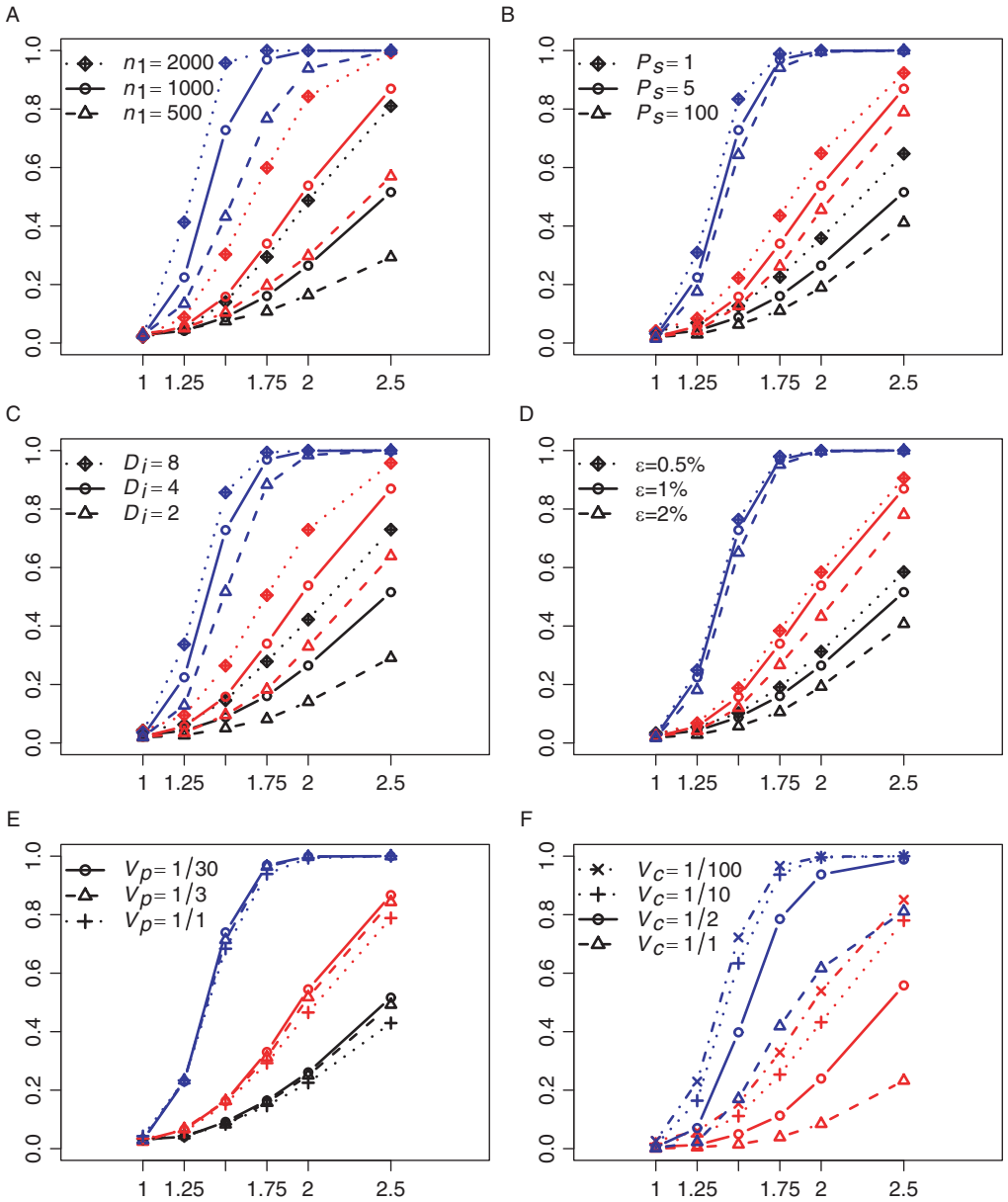

Fig. 3. Power comparison across a range of sample sizes (number of cases, n_1), pool sizes P_s, individual sequencing depths D_i, sequencing error rates ε, pooling variances V_p, and capturing variances V_c. Note that the power is defined as the probability of including the causative SNP among the set of the 50 SNPs with strongest signal of association among 300,000 sites. Each plot shows the power (y-axis) to detect a causative SNP with a specified minor allele frequency (black: 0.5%; red: 1%; and blue: 5%) and a specified relative risk (x-axis), across a range of a parameter settings of interest (line types). The default experimental setting is to sequence 1000 cases and 1000 controls at 4× using a pool size of five, and the default condition assumes a sequencing error rate of 1%, ideal pooling performance, and constant capturing efficacy. Notice that the solid line in the plots A–D corresponds to the default setting and condition.

reasonably accurately estimated as long as the pool size is not too large (i.e., >100), and thus the likelihood ratio test achieves a power very close to the case when the true error rate is known. Additionally, most sequencing platforms provide quality scores that can be interpreted as error rates. Even when these quality scores are not quite accurate, they can be re-calibrated to accurately reflect the error rate. Mis-specification of the error rate is, therefore, not likely to affect carefully constructed real studies.

COMPARISON WITH *G*-TEST

We have compared the power of the LRT statistic to a *G*-test across the range of experimental settings presented in Figure 3. As expected, we find overall that our statistic

performs better than the *G*-test (Supplementary Fig. 9), especially when analyzing rare disease variants (MAF < 5%). The *G*-test does not take the pooling structure into account, and therefore, the power of the test is not improved much with a decreased pool size (Supplementary Fig. 10B). Also, the *G*-test does not account for the feature of next-generation sequencing data, in which, multiple copies are potentially produced from a single allele. Indeed, ignoring this feature causes the distribution of the *G*-statistics to deviate from the expected asymptotic $\chi^2(1)$ distribution (Supplementary Fig. 11). The deviations from the χ^2 distribution is especially large when the allele-frequencies and individual sequencing depth are high. Interestingly, due to this property, when the individual sequencing depth is very high, the test statistic separates true SNPs from false SNPs very well. For false SNPs, every minor-allele read is independently generated due to sequencing errors, and therefore, the test statistic approximately follows a χ^2 distribution. However, statistics calculated on true SNPs tend to be large, and so, any SNP including the disease SNP is likely to be among the ranked sites.

To evaluate the performance of the statistics in terms of efficacy in detecting the causal SNPs among other SNPs, we also compared the two statistics when the null set consists of only segregating sites (true SNPs). The average rank of the disease SNP in 5,000 repeated simulations is lower for the LRT statistic than for the *G*-statistic (Supplementary Fig. 12), again implying that the correctly specified likelihood ratio statistic is better at distinguishing allele frequency differences between cases and controls.

OPTIMAL DESIGN WITH A FIXED COST

In this section, we explore options for maximizing the power for a fixed experimental cost. (Examples of cost are shown in Table I.) For a given cost, the factors we can control are the number of cases n_1, pool size P_s, and the individual sequencing depth D_i. Note, however, that the power is also affected by the sequencing error rate, pooling performance, and exon-capturing effects.

UN-POOLED SAMPLES

Even if pooling DNA samples may reduce cost significantly, sequencing un-pooled samples is of great interest as well, since eventually, sequencing whole genome might be feasible at a low cost. An important question then arises as how to balance the trade-off between the number of individuals and the individual sequencing depth, with a fixed sequencing cost. To illuminate this issue, we examined three designs with 500 cases, 1,000 cases, and 2,000 cases at a fixed total depth of 8,000 ×. Under our simulation conditions, power increases as the number of individuals increases (Fig. 4). For example, when the error rate is 1%, for a disease variant with MAF of 1% and a relative risk of 2, the power using 16 × sequencing of 500 individuals is 52.1%, while the power using 4 × sequencing of 2,000 individuals is 91.1%. When the total depth is reduced by half, so that the individual depth for 2,000 cases is 2 ×, the pattern is still the same (Supplementary Fig. 13). These observations

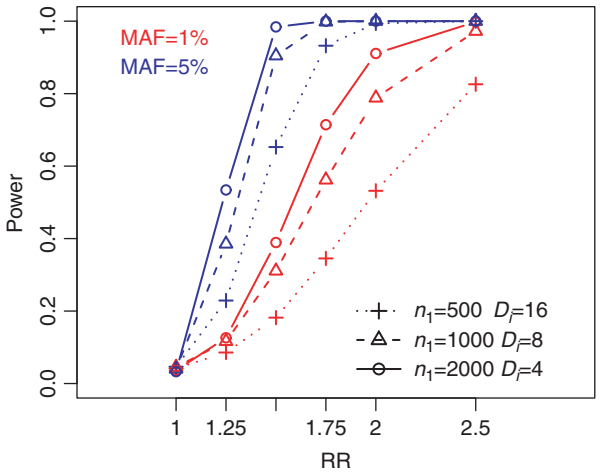

Fig. 4. Power comparison among different sequencing strategies with a fixed sequencing cost. Power is defined as in Figure 3. The different strategies are specified in the bottom right corner of each plot. n_1 and D_i specifies the number of cases and sequencing depth per individual, respectively. Notice that the total sequencing depth is fixed at 8,000 × across the three settings compared. The power to detect a causative SNP with a minor allele frequency (red: 1%; blue: 5%) and a specified relative risk (*x*-axis) are compared for different settings (line types).

suggest that with a reasonable sequencing error rate, such as 1%, individual sequencing of many individuals at a shallower depth achieves better power than sequencing less individuals at a higher depth. This finding might be surprising given that it is difficult to call rare SNPs in the presence of high error rates. Real error rates might be lower than the 1% assumed here, which would further strengthen our conclusions. If DNA from many individuals is readily available, there seems to be very little reason to sequence few individuals at a depth > 2 ×. Rather, statistical power is maximized by sequencing more individuals at a lower depth (e.g., 2 ×).

POOLED SAMPLES

We investigated optimal strategies for pooled samples with a fixed cost. To fix the cost, we fix the total number of pools as well as the total depth, thus fixing exon-capturing cost as well as sequencing cost. We use four different parameter settings, in which, the setting with the lowest sample size corresponds to sequencing 500 cases individually at 8 ×. The setting with the biggest sample size corresponds to sequencing 4,000 cases at 1 × with a pool size of 8 individuals. The four settings are shown in the *x*-axis of Figure 5A, in the order of sample size. For each parameter setting, we compared the power to detect a disease SNP with MAF of 1% and RR of 2, under various experimental scenarios, considering the error rate, pooling performance as well as exon-capturing efficacy. With an error rate of 1%, the power increases with sample size, even in the presence of a more shallow sequencing depth and a larger pool size (Fig. 5A). This may be surprising since, when individuals are pooled, it becomes even harder to distinguish true SNPs from errors. With a lower error rate of 0.5%, the gain in power with increased sample

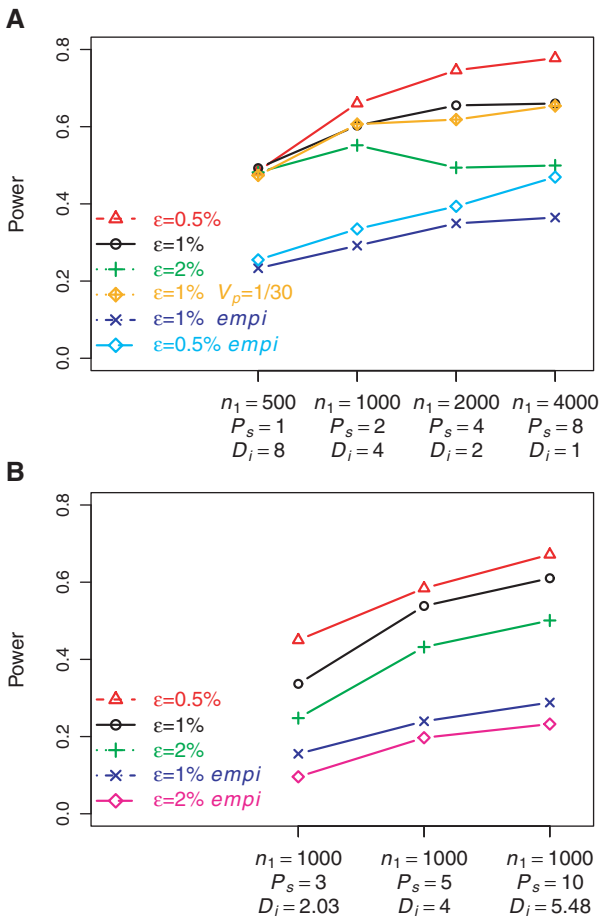

Fig. 5. Optimal strategies with pooled samples. The power to detect a causative SNP with MAF of 1% and relative risk of 2 is compared for different experimental settings with a fixed cost. Power is defined as in Figure 3. In (A), the total number of pools and the total depth are fixed across the four settings, but the sample size is varied (*x*-axis). In (B), the sample size is fixed across the three settings, but the depth varies (*x*-axis). Each symbol corresponds to different sequencing error rates, pooling variance, and exon-capturing variance. In the legend, ε specifies an error rate pooling variance is specified by V_p and *empi* denotes the condition using the pooling variance and exon-capturing variance estimated from the empirical study (see Results).

size is more pronounced. For example, when the error rate decreases from 1 to 0.5%, the power of sequencing 500 cases individually at 8 × remains almost the same. While the power of sequencing 4,000 cases at 1 × with a pool size of 8 increases from 66 to 78%. This shows that when the individual depth is sufficient, not much is gained by having a low error rate, but when depth is shallow and pool size is relatively large, then the low error rate helps considerably in distinguishing rare SNPs from errors. For similar reasons, a very high error rate such as 2% tends to make designs with a shallow sequencing depth and large pool size less favorable. Indeed, with an error rate of 2%, we observe that the power when sequencing 2,000 individuals at 2 × with a pool size of 4 is lower than when sequencing 1,000 individuals at 4 × with a pool size of two. However, such high error rates are unlikely to occur in practice. Assuming the pooling variance and exon-capturing variance estimated in the empirical study, the power for all the settings decreases but the overall pattern remains the same. The decrease in the power is mostly due to variation in exon-capturing efficacy, since introducing pooling variance only does not result in a significant loss in the power (Fig. 5A). We also note that the overall pattern discussed above holds true for a range of MAFs and RR, especially for the most relevant ones, such as a MAF higher than 1% and RR larger than 1.3.

As the number of individuals available for sequencing or genotyping often is fixed, it is also of relevance to consider designs in which both the cost and the number of individuals is fixed, but the sequencing depth and pool size can vary. We create three experimental settings based on sequencing 1,000 individuals. Figure 5B shows that with a fixed number of individuals, it is better to sequence at a higher depth with a larger pool size than to sequence in a more shallower depth with a smaller pool size. For example, with a sequencing error rate of 1%, a pool size of 3 and a sequencing depth of 2 ×, the power is 45%. However, with a pool size of 10 and a sequencing depth of 5.48 × the power is 67.2%. A similar pattern is observed across a range of sequencing error rates from 0.5 to 2%. This conclusion also holds when taking pooling variance and exon-capturing variance into account (Fig. 5B).

Overall, our finding suggests that sequencing pooled DNA samples in the region of interest works as an efficient protocol in association studies, especially when there is a significant cost involved in capturing the targeted regions. In particular, when the cost of obtaining DNA from individuals is low, sequencing many individuals at a more shallow depth with larger pool size achieves higher power than sequencing a small number of individuals in higher depth with smaller pool size. When a fixed number of individuals are available for the study, it is better to pool more individuals to save cost in capturing rather to spend more money on sequencing each pool deeper.

DISCUSSION

We have presented here a simple likelihood ratio test for association mapping based on next-generation sequencing data in pooled or un-pooled samples. The method can be improved in a number of ways, to take into account quality scores and known error rates. The real pattern of sequencing errors is much more complicated than assumed in our simulations [e.g., Li et al., 2009]. Most of these potential extensions are relatively trivial and should not affect the conclusions from the simulation studies presented here. However, it might be worthwhile in future studies to explore other methods than independent likelihood ratio tests for detecting associated SNPs. It might be natural to implement empirical Bayes approaches, which first estimate the number of SNPs with an association and then provide a posterior for each SNP. Such methods could also attempt to take advantage of LD and haplotype patterns in the data. In many real studies, it might also be relevant to use imputation to leverage available SNP data from other studies [Marchini et al., 2007; Scott et al., 2007; Servin and Stephens, 2007]. However, we have not pursued such approaches here as the primary aim of this

paper is to use very fast computational methods to explore issues regarding design of association mapping studies based next-generation sequencing data.

We used a small empirical study to estimate parameters relating to exon-capturing variance and pooling variance in order to have reasonable parameters for use in the simulations. In general, none of the conclusions in the manuscript seems particularly sensitive to assumptions regarding exon-capturing and pooling variance. However, we should note that our conclusions may not necessarily generalize to pools much larger than 5 individuals. Our assumption has been that pooling variance does not depend on sample size. If it is considerably more difficult to make pools of, say, 50 individuals than 5 individuals, this assumption may not hold. We do not have sufficient data to determine the potential accuracy of pooling in large samples.

We used the likelihood ratio test and the empirically estimated parameters to simulate association mapping studies under realistic conditions. In these simulations we ignored LD to increase the computational speed. While this is not a realistic assumption, it should not affect our general results, as LD is not expected to have a different effect on pooled and un-poled samples, or in samples with deep versus shallow sequencing, and since our tests statistic is based on marginal analyses of each SNP.

Our major conclusions from the simulation studies is that relatively shallow sequencing and relatively large pools are almost always preferred. We have assumed quite high error rates in this study, in part because higher error rates make identification of rare mutations harder, and this effect is exasperated under shallow sequencing and in pools where reads from each individual cannot be identified. Assuming high error rates, therefore, biases our conclusions in the opposite direction of the observed results, toward deep sequencing and small pools. It might be surprising that even when rare mutations are difficult to identify, shallow sequencing is still preferred. The main reason is that the power to detect an association depends on the ability to estimate allele frequencies among cases and controls. With a fixed amount of sequencing data being produced, the variance in the estimate of the allele frequency is always lower with more individuals than with few individuals. This effect is balanced by a potentially reduced ability to determine which sites contain true SNPs and which sites do not contain true SNPs. However, as illustrated by the simulations, any possible reduction in our ability to call SNPs does not offset the gains in power achieved by the more accurate estimation of allele frequencies.

As sequencing cost continues to decrease, and as focus is changing toward rare alleles, next-generation sequencing is becoming a viable alternative for association mapping. We emphasize that at the moment, economically optimal designs are probably not based entirely on next-generation sequencing. A two-stage approach involving sequencing in the first stage and genotyping in the second stage is almost always a more efficient design, especially since imputation can be used in the second stage to increase power [Marchini et al., 2007; Scott et al., 2007; Servin and Stephens, 2007]. In fact, power in our studies was measured in terms of the probability of selecting a causative SNP in the first stage of a two-stage design. However, it is possible that in the near future, sequencing becomes so cheap that designs based entirely on sequencing should be considered. Our conclusions favoring shallow sequencing in many individuals and pooled samples as a cost-reducing measure should be relevant in both candidate gene studies using sequencing and in GWAs based on either a mixed two-stage design or based entirely on next-generation sequencing.

REFERENCES

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA. 2007. Direct selection of human genomic loci by microarray hybridization. Nat Methods 4:903–905.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. Science 322:881–888.

Azzopardi D, Dallosso AR, Eliason K, Hendrickson BC, Jones N, Rawstorne E, Colley J, Moskvina V, Frye C, Sampson JR, Wenstrup R, Scholl T, Cheadle JP. 2008. Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res 68:358–363.

Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A. 2002. Association testing by DNA pooling: an effective initial screen. Proc Natl Acad Sci USA 99:16871–16874.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701.

Boss Y, Bacot F, Montpetit A, Rung J, Qu HQ, Engert JC, Polychronakos C, Hudson TJ, Froguel P, Sladek R, Desrosiers M. 2009. Identification of susceptibility genes for complex diseases using pooling-based genome-wide association scans. Hum Genet 125:305–318.

Broyden CG. 1970. The convergence of a class of double-rank minimization algorithms. J Inst Maths Applics 6:222–231.

Chi A, Schymick JC, Restagno G, Scholz SW, Lombardo F, Lai SL, Mora G, Fung HC, Britton A, Arepalli S, Gibbs JR, Nalls M, Berger S, Kwee LC, Oddone EZ, Ding J, Crews C, Rafferty I, Washecka N, Hernandez D, Ferrucci L, Bandinelli S, Guralnik J, Macciardi F, Torri F, Lupoli S, Chanock SJ, Thomas G, Hunter DJ, Gieger C, Wichmann HE, Calvo A, Mutani R, Battistini S, Giannini F, Caponnetto C, Mancardi GL, La Bella V, Valentino F, Monsurr MR, Tedeschi G, Marinou K, Sabatelli M, Conte A, Mandrioli J, Sola P, Salvi F, Bartolomei I, Siciliano G, Carlesi C, Orrell RW, Talbot K, Simmons Z, Connor J, Pioro EP, Dunkley T, Stephan DA, Kasperaviciute D, Fisher EM, Jabonka S, Sendtner M, Beck M, Bruijn L, Rothstein J, Schmidt S, Singleton A, Hardy J, Traynor BJ. 2009. A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. Hum Mol Genet 18:1524–1532.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872.

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH. 2006. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci USA 103:1810–1815.

Druley TE, Vallania FL, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, Fay JC, Mitra RD. 2009. Quantification of rare allelic variants from pooled genomic DNA. Nat Methods 6:263–265.

Fletcher R. 1970. A new approach to variable metric algorithms. Comp J 13:317–322.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27:182–189.

Goldfarb D. 1970. A Family of variable-metric methods derived by variational means. Math Comp 24:23–26.

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardisom ER. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods 5:183–188.

Hindorff LA, Junkins HA, Mehta JP, Manolio TA (Accessed on April 1, 2009). A Catalog of Published Genome-Wide Association Studies. *National Human Genome Research Institute*. Available at www.genome.gov/26525384.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. 2007. Genome-wide in situ exon capture for selective resequencing. Nat Genet 39:1522–1527.

Ingman M, Gyllensten U. 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. Eur J Hum Genet 17:383–386.

Iyengar SK, Elston RC. 2007. The genetic basis of complex traits: rare variants or ''common gene, common disease''? Methods Mol Biol 376:71–84.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP. 2008. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40:592–599.

Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC, Palmisano E, Brune K, Jaffee EM, Iacobuzio-Donahue CA, Maitra A, Parmigiani G, Kern SE, Velculescu VE, Kinzler KW, Vogelstein B, Eshleman JR, Goggins M, Klein AP. 2009. Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. Science 324:217.

Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R. 2008. A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc Natl Acad Sci USA 105:9296–9301.

Kruglyak L. 2008. The road to genome-wide association studies. Nat Rev Genet 9:314–318.

Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. 2009. Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106:3871–3876.

Lavebratt C, Sengul S. 2006. Single nucleotide polymorphism (SNP) allele frequency estimation in DNA pools using Pyrosequencing. Nat Protoc 1:2573–2582.

Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009. SNP detection for massively parallel whole-genome resequencing. Genome Res 19:1124–1132.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913.

McCarthy MI, Hirschhorn JN. 2008. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:R156–R165.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369.

McClellan JM, Susser E, King MC. 2007. Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry 190:194–199.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324:387–389.

Nelder JA, Mead R. 1965. A simplex method for function minimization. Comp J 7:303–307.

Norton N, Williams NM, O'Donovan MC, Owen MJ. 2004. DNA pooling as a tool for large-scale association studies in complex traits. Ann Med 36:146–152.

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. 2007. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4:907–909.

Polychronakos C. 2008. Common and rare alleles as causes of complex phenotypes. Curr Atheroscler Rep 10:194–200.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J. 2007.

Multiplex amplification of large sets of human exons. Nat Methods 4:931–936.

Raymond FL, Whibley A, Stratton MR, Gecz J. 2009. Lessons learnt from large-scale exon re-sequencing of the X chromosome. Hum Mol Genet 18:R60–R64.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341–1345.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114.

Sham P, Bader JS, Craig I, O'Donovan M, Owen M. 2002. DNA pooling: a tool for large-scale association studies. Nat Rev Genet 3:862–871.

Shanno DF. 1970. Conditioning of Quasi-Newton methods for function minimization. Math Comp 24:647–656.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213.

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF, Hoover RN, Chanock SJ, Hunter DJ. 2009. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41:579–584.

Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods 5:247–252.

Wang H, Thomas DC, Pe'er I, Stram DO. 2006. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 30:356–368.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. 2008. The diploid genome sequence of an Asian individual. Nature 456:60–65.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876.

Zuo Y, Zou G, Zhao H. 2006. Two-stage designs in case-control association analysis. Genetics 173:1747–1760.

Zuo Y, Zou G, Wang J, Zhao H, Liang H. 2008. Optimal two-stage design for case-control association analysis incorporating genotyping errors. Ann Hum Genet 72:375–387.