Published in final edited form as:

Genet Epidemiol. 2012 December ; 36(8): 870–881. doi:10.1002/gepi.21681.

Analysis and Optimal Design for Association Studies Using Next-Generation Sequencing With Case-Control Pools

Wei E. Liang, Duncan C. Thomas, and David V. Conti*

Department of Preventive Medicine, University of Southern California, Los Angeles, California

Abstract

With its potential to discover a much greater amount of genetic variation, next-generation sequencing is fast becoming an emergent tool for genetic association studies. However, the cost of sequencing all individuals in a large-scale population study is still high in comparison to most alternative genotyping options. While the ability to identify individual-level data is lost (without bar-coding), sequencing pooled samples can substantially lower costs without compromising the power to detect significant associations. We propose a hierarchical Bayesian model that estimates the association of each variant using pools of cases and controls, accounting for the variation in read depth across pools and sequencing error. To investigate the performance of our method across a range of number of pools, number of individuals within each pool, and average coverage, we undertook extensive simulations varying effect sizes, minor allele frequencies, and sequencing error rates. In general, the number of pools and pool size have dramatic effects on power while the total depth of coverage per pool has only a moderate impact. This information can guide the selection of a study design that maximizes power subject to cost, sample size, or other laboratory constraints. We provide an R package (hiPOD: hierarchical Pooled Optimal Design) to find the optimal design, allowing the user to specify a cost function, cost, and sample size limitations, and distributions of effect size, minor allele frequency, and sequencing error rate.

Keywords

genetic association studies; sequencing; rare variants

INTRODUCTION

The last decade has seen a rapid growth of genome-wide association studies (GWAS). GWAS look for the association between genetic variation (mainly single-nucleotide polymorphisms, SNPs) and phenotypes of interest in a hypothesis-free manner. These studies were so popular that around 6,000 SNP associations have been reported for more than 130 traits as of January 2012 (<http://www.genome.gov/gwastudies/>). However, only a small-to-moderate proportion of the hereditary variation has been explained by the variants

discovered [Frazer et al., 2009]. To find the unexplained heritability by GWAS, investigators have begun to look beyond common genetic variants. Based on population genetics, deleterious variants are often under greater selection pressure, and they are more likely to be present in a lower frequency in the current population. There have already been a few published studies indicating the effects of rare variants, solely or in combination, in common diseases [Li and Leal, 2008; Morris and Zeggini, 2009; Neale et al., 2011; Quintana et al., 2011].

To better study the effects of rare variants in association studies, technological advances, especially next-generation sequencing (NGS), have been rapidly developing. Despite its efficiency for generating sequencing reads, NGS is still relatively expensive to be used in population-based association studies. The structure of the NGS data is also completely different from current genotyping technologies, and the error rate of NGS is relatively higher than current genotyping technologies [Shendure and Ji, 2008]. Aiming for uncompromised statistical power with much reduced cost, DNA pooling has become a promising design for association studies, especially applicable to NGS data [Lupton et al., 2011; Out et al., 2009; Prabhu and Pe'er, 2009; Wang et al., 2011].

Prabhu and Pe'er [2009] constructed over-lapping pools to tackle the problem of detecting the carrier of rare variants. Similar works along these lines are Shental et al. [2010] and Erlich et al. [2009]. Price et al. [2010] also carried out tests for detecting rare variants associations from pools. Druley et al. [2009] developed an efficient algorithm to accurately quantify rare variants from pooled genomic DNA using NGS. Futschik and Schlotterer investigated some statistical properties of methods for detection and frequency estimation of rare variants in NGS data of pooled samples [Futschik and Schlotterer, 2010]. Lee et al. [2011] tried to find the optimal number of individuals in a DNA pool to detect an allele under given coverage depth and detection threshold. Focusing on association studies, Kim et al. [2010] developed a likelihood ratio test on pooled or un-pooled designs, and investigated the design that achieved the optimal power. They elegantly considered the structure of the NGS data, especially the nature of the NGS error. However, there are still some essential parts missing for a thorough analysis of such a design: a clear criterion to discriminate SNPs from sequencing error, a thorough examination of the validity of the study design, robust estimation of SNP frequency and effect size, a realistic cost function reflecting the different costs of pool formation and sequencing, and corrections for multiple tests.

Here, we propose a Bayesian framework for addressing these issues. Capturing the exact NGS data structure, we show that our method yields an unbiased estimation not only of the variant frequency in a pool, but also of the effect size of a variant for a robust test on the significance of the association. Using extensive simulations, we have described the constraints for valid study designs that yield the expected test statistic distribution under the null. Among valid designs, we show that our test statistic has a normal distribution and that the scale parameter of the test statistic was well predicted by our design parameters. Based on these results, we developed a formula for power calculations incorporating all the design parameters and the assumed genetic model, and extended this approach to the GWAS setting. Using a realistic model for the dependence of NGS costs on the design parameters, we have developed a statistical tool for finding an optimal design subject to the budget

limitations, the *VAF*, the effect size of the SNP to be targeted, and the error rate of the NGS instrument. For researchers looking for a cost-efficient association study design, we hope our work will provide useful insights on the design and analysis of such studies in pooled samples with NGS technology.

METHODS

A HIERARCHICAL BAYESIAN MODEL FOR POOLED SEQUENCING DATA

In a nonoverlapping pooled sequencing study, individuals are randomly allocated into pools within each stratum (e.g., case/control status). At each location, we observe the total number of reads X_p and the number of variant reads V_p in pool p . We model V_p as a Binomial distribution with probability θ_p .

(1)

where λ is the average depth of sequencing for each pool and $\gamma_p \sim \text{Gamma}(\alpha, 1/\alpha)$ accounts for variation in DNA-capturing efficacy; note that this implies that the average depth of sequencing per subject is λ/N_p . V_p/X_p may be used as an estimate of the variant frequency; however, the standard error of this estimate will be greatly underestimated as it is driven by the coverage and not the number of individuals within each pool. Since pool p constitutes DNA samples from N_p individuals, we let Z_p denote the number of chromosomes (out of the $2N_p$ possible chromosomes) that carry the variant allele in pool p . Thus, ρ_p denotes the variant allele frequency for the individuals in pool p , and Z_p is distributed as:

(2)

Since each read is generated as an independent Bernoulli process, in the absence of sequencing error, the probability of a read being called as a variant is $Z_p/2N_p$. To account for sequencing error, we assume a simple and symmetric error structure in which $\varepsilon = \Pr(\text{read} = v | \text{allele} = r) = \Pr(\text{read} = r | \text{allele} = v)$, where v denotes the variant allele and r denotes the reference allele. Thus, in the presence of sequencing error, θ_p is defined as:

(3)

To facilitate case-control comparisons across pools, we define the variable W_p as an indicator of the disease status of all individuals within pool p ($W_p = 1$ if pool p constitutes cases and $W_p = 0$ if pool p constitutes controls), and we use a logistic model to link the pool-level variant allele frequency with the case-control status of the pool:

(4)

Here, β_0 is the logit of the variant frequency within the entire sample; β_1 is the log of the odds ratio between cases and controls used for association testing (e.g., with disease or determinants of the *VAF*); and $\bar{W}$ is the average of W . For completeness, we tried different prior distributions on the hyperparameters, and found that the model is insensitive with respect to the prior distribution, especially in valid study designs. Thus, we put relatively noninformative priors on hyperparameters, and fit the entire model using Markov chain Monte Carlo (MCMC) techniques as implemented in WinBUGS.

(5)

Overall, our hierarchical model is summarized in the DAG (Directed Acyclic Graph) in Figure 1.

SIMULATION STUDIES

To investigate the performance of our hierarchical model and to build a prediction model for power, we performed simulations over a variety of scenarios with different variant allele frequencies (*VAF*), effect sizes (*OR*), and sequencing error rates (ϵ), as well as a range of study design parameters including the number of pools (P , thus there are $P/2$ case pools and $P/2$ control pools), the number of individuals per pool (N_p), and the average coverage for each pool (λ).

Given specified values for the samples *VAF*, *OR*, and ϵ , we simulated the number of variant-carrying chromosomes in each pool using equations (2) and (4)). Within each pool, we simulated the total coverage X_p and the number of variants V_p following equations (1) and (3).

The simulations used the following specified values: *VAF* = {0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.12, 0.15, 0.2}, *OR* = {1.0, 1.2, 1.6, 2.0}, and ϵ = {0.005, 0.01, 0.02}. We varied the design parameters with N_p = {1, 2, 5, 20, 40, 60, 80, 100, 150, 200}; P = {2, 4, 10, 40, 80, 120, 160, 200, 250, 300, 350, 400}; and λ = {10, 20, 40, 80, 160, 320, 640, 1,280}. For each specified scenario, we simulated 1,000 replicates and analyzed the resulting data with our hierarchical model using WinBUGS and a single chain of 1,000 MCMC samples. We summarized the posterior distribution of β_0 , β_1 , and ϵ . Hypothesis testing was based on the test statistic $B = \hat{\beta}_1 / \hat{\sigma}_1$ for each replicate, where $\hat{\beta}_1$ and $\hat{\sigma}_1$ are the posterior mean and standard error of the effect size. Under the null hypothesis, we assume that B is distributed as a standard normal. The empirical power was calculated as the proportion of replicates that show a statistical significance to reject the null hypothesis (P -value < 0.05), and empirical type I error was the empirical power under the null scenarios (where the simulated *OR* = 1).

DETERMINING OPTIMAL STUDY DESIGN FOR FUTURE STUDIES

Using Simulations to Build a Prediction Model for Power

In valid designs, the test statistics should be distributed as a standard normal under a null scenario. Thus, we used a Kolmogorov-Smirnov (K-S) test to determine whether these test statistics obtained from the hierarchical model under a null scenario for a given study design follow a standard normal distribution, which is a prerequisite for the study design to be valid. If the corresponding K-S test was significant (P -value < 0.05), we removed the study design from the determination of the prediction model. Thus, in all the valid designs, the test statistic follows an asymptotically normal distribution summarized by the location parameter of each valid design, denoted as μ_d with $B_d \sim N(\mu_d, 1)$. We empirically estimated the

location parameter μ_d using B_{dr} , where R is the number of replicates for design d , and B_{dr} is the test statistic for design d replicate r . Using these estimated location parameters, we then built a linear regression model to predict μ_d as a function of all the scenario parameters (VAF , OR , and ε) and design parameters (N_p , P , and λ). This regression model is then used to predict power for future study designs.

Predicting Power for Future Studies

To predict power for any proposed future study, we first determined if the study design results in valid inference using our hierarchical modeling analysis, and if deemed valid, we then predicted power using the estimated prediction model constructed from the analysis of the simulated results.

For the examples in this paper, we determined if a proposed study was valid by calculating the probability that a variant will be detected in either the cases or the controls. For a specific design, we use a threshold for the number of variant reads, V_t to determine if a variant is detected. Given the pool size (N_p), average coverage per pool (λ), and number of pools (P), if there was no variant in the pool, the number of variant reads (V_0) would be approximately distributed as: $V_0 \sim \text{Bin}(\lambda, \theta_0 = \varepsilon)$, where we use the average coverage λ as an approximate of the true coverage. If only one chromosome in the pool carries a variant, the number of variant reads (V_1) would be approximately distributed as: $V_1 \sim \text{Bin}(\lambda, \theta_1 = (1 - \varepsilon)/2N + \varepsilon(2N - 1)/2N)$. By comparing these two situations, we calculate two corresponding probabilities: $\Pr(V_0 > V_t)$ and $\Pr(V_1 > V_t)$. To control type I error with Bonferroni correction for multiple tests at a significance level of 0.05 and to ensure an adequate odds that a true variant is detected vs. a false discovery, we chose the smallest V_t that satisfies both $\Pr(V_0 > V_t) < (0.1/P)$ and $\Pr(V_1 > V_t) > 5\Pr(V_0 > V_t)$. This is a conservative comparison, since $\Pr(V_1 > V_t) < \Pr(V_2 > V_t) < \dots < \Pr(V_{2N} > V_t)$, and we are comparing the least distinctive positive scenario vs. negative scenario.

Next, we used the threshold V_t to calculate the probability of variant detection. If we assume that we have an equal number of case and control pools, the probability of detecting a true variant in both the case pools and the control pools is:

Valid study designs were those in which $\Pr(\text{detection}) > \Pr_T$, where $\Pr_T$ is determined from a comparison of the validity determination using the K-S test to the predicted validity using the calculated $\Pr(\text{detection})$ for all simulated scenarios.

If a novel study design was determined to be valid, then we predicted the expected noncentrality parameter using the fitted regression model. Let μ_d denote the predicted noncentrality parameter, P_d denote the number of pools in the proposed study design, and q_{ad} denote the one-sided α -level critical value for the test statistic: $\Pr[x > q_{ad} | x \sim N(0, 1)] = \alpha$. The expected power for the proposed study design is $\Lambda_d = \Pr[x > q_{ad} | x \sim N(\mu_d, 1)]$. Using the estimated noncentrality parameter $\hat{\mu}_d$ from the prediction model, we predicted the expected power as $\Lambda_d = \Pr[x > q_{ad} | x \sim N(\hat{\mu}_d, 1)]$.

We obtained the 95% confidence interval for the noncentrality parameter μ_d , denoted $\hat{\mu}_{ld}$ and $\hat{\mu}_{ud}$, thus $\Pr(\mu_d > \hat{\mu}_{ud}) = \Pr(\mu_d < \hat{\mu}_{ld}) = 0.025$. The confidence interval for the predicted power was based on the confidence interval of the noncentrality parameter.

To test the prediction model, we also performed an empirically based simulation study using the 1000 Genomes exome alignment data (based on the exome alignment index file in <ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/exome.alignment.index>), including unrelated individuals in both pilot project 2 and pilot project 3. Focused on a subset of representative SNPs in Chromosome 2 (as shown in Table I), we extracted the sequencing reads and genotype of each individual, and simulated the disease status based on a logistic disease model. Based on the specific study design (i.e., number of pools $P = \{20, 40, 100, 200\}$, number of individuals per pool $N_p = 1,000/P$, and average coverage per pool $\lambda = \{20, 50\}$), we randomly sampled affected/unaffected individuals to form case/control pools, and constructed the pool-level sequencing data by summing over a randomly sampled subset of sequencing reads of each member of the pool. Finally, we applied our Bayesian hierarchical model to analyze the generated pooling association study data. For each scenario, we generated 1,000 replicates and compared the empirical power to the power calculated from our prediction function.

Finding an Optimal Study Design

Given a set of assumed scenario parameters (VAF , OR , and ε), we used the following procedure to find an optimal study design:

1. Specify the range of interest on design parameters (P , N_p , and λ) with constraints on study cost (C_T) and sample size (N_T).
2. Using a generalized cost function:

where C_E is the cost per experiment including the initiation of the study and sample collection, C_P is the cost per pool such as sample preparation (and region capturing if focused on region specific sequencing), and C_X is the cost per $1\times$ coverage.

3. Specify 120×120 grid covering values of P and N_p across the specified range of interest.
4. For each point in the grid use the corresponding number of pools (P_g) and number of individuals per pool (N_{pg}) to calculate the maximum coverage λ_g consistent with the constraint to study cost (C_T).
5. Check if the design (P_g, N_g) satisfies all of the following conditions:
 - a. $P_g N_g \leq N_T$ to satisfy the sample size constraint;
 - b. Calculate the $\text{Pr}(\text{detection})$ for the specific design using the maximum coverage λ_g . $\text{Pr}(\text{detection}) \geq \text{Pr}_T$ are deemed valid study designs.
 - c. Using the estimated prediction model, calculate the predicted power for the study design.
6. Perform an exhaustive search over all valid grid-points. Select the optimal design as the study with the maximum power.

To help investigators find the optimal pooling design, an R package hiPOD (hierarchical Pooled Optimal Design) is now available in the online CRAN depository.

RESULTS

VALIDITY AND EMPIRICAL POWER OF SIMULATED SCENARIOS

In total, we simulated 116,160 scenarios with different study designs and disease models. Not all the scenarios result in valid tests of association. Results from the K-S tests for the uniform distribution of type I error under the null scenarios demonstrated that, in general, study designs with few pools, few individuals within each pool, or with low coverage tend to result in invalid tests of association. This was especially true for scenarios dealing with rare variants and small effect sizes. These trends were reflected in Figure 2, showing the valid and invalid designs for a variant with $VAF = 0.05$ and an average coverage of either 40 or 160, respectively. The reason large N_p designs are often invalid in Figure 2 is that the average depth of sequencing *per subject* is very low for some of these.

From all the simulated scenarios, we selected a few representative examples to illustrate the relationship between empirical power and the design parameters. In Figure 3A and B, the variant allele frequency is 0.05 with a moderate effect size (1.6). In such a setting, empirical power increases as either coverage increases or more individuals are included in each pool. In the figure also displays the optimal power from a comparable individual sequencing study, which is assumed to have no errors in genotyping calls due to high coverage, and have the same number of individuals sequenced as the number of pools constructed in the pooling design (thus they would require similar sequencing efforts). Such a separately constructed individual sequencing analysis has an improved power over the single-individual pool analysis, as it eliminates the within-pool uncertainties in the analysis. In this

scenario, the pooling design has superior power over individual sequencing, especially as the number of individuals in each pool increases or as the coverage increases. Figure 3C–E show a scenario where a rare variant ($VAF = 0.01$) with a large effect ($OR = 2.0$) is tested. Similar results are obtained with increased empirical power as the number of pools increases, the coverage increases or as the number of individuals in each pool increases. In all the plots of Figure 3, we include both valid and invalid designs to demonstrate overall trends in power. However, while it is true that increasing the number of individuals per pool increases power, if the average coverage per pool is held constant, the study design is more likely to yield an invalid test.

In general, power increases most with number of pools, whereas number of individuals per pool has a moderate effect and the coverage depth per pool has only a limited effect. These trends are more clearly seen in contour plots of power with respect to the design parameters (Figure 4).

CONSTRUCTION OF A PREDICTION MODEL

To predict power for future study designs, we removed all invalid designs and all null scenarios from the simulated data. We confirmed that the test statistic was asymptotically distributed as normal, empirically summarizing location parameter μ_d with B_d . We subsequently used B_d to predict power and found that the predicted power compared very favorably to the empirical power observed, with an $R^2 = 0.998$.

In constructing prediction models for the noncentrality parameter of the test statistics, we built separate prediction models for rare variants ($VAF < 0.01$), less-common variants ($0.01 < VAF < 0.05$), modest frequency variants ($0.05 < VAF < 0.1$), and common variants ($VAF > 0.1$). We transformed the response variable for better model prediction. We used a linear regression model for the cube root of the noncentrality parameter as a function of the variant allele frequency, log odds ratio, sequencing error rate, number of pools, number of subjects per pool, average depth of sequencing, and the various transformations and interactions among these variables listed in Table II. In general, all the four models (for different strata of VAF s) predict the noncentrality parameter quite accurately (all with $R^2 > 0.97$). We summarized the partial R^2 for each variable in Table I for the full model, bolding the two strongest predictors after $\ln(OR)$. Across all the design parameters, the average sequencing coverage depth explains the most variation in performance involving rare VAF s (VAF 0.002–0.01), whereas the number of pools explains the greatest variations involving intermediate and common VAF s (VAF 0.05–0.2). The predicted power is highly correlated ($R^2 = 0.97$) with the empirical power.

PREDICTING POWER AND DETERMINING OPTIMAL DESIGNS FOR FUTURE STUDIES

For future studies with a given design in terms of number of pools, number of individuals within each pool, and coverage, we calculate the probability of variant detection and use this value to decide if the proposed study is valid. To determine the variant probability cutoff, we use the simulated data scenarios and results to construct an ROC (Receiver Operating Characteristic) curve comparing these calculated probabilities to the validity determination from the K-S test, using the K-S test as the gold standard. Overall, the area under the ROC

curve was equal to 0.89. From this curve, we selected a cutoff at 0.9, yielding a sensitivity of 0.91 and specificity 0.74.

We used the estimated prediction models to calculate power for any proposed future study design. For each valid study design, power prediction involves two steps: first predict the noncentrality parameter for the test statistic, and then calculate power based on the predicted noncentrality parameter. In the simulation study based on 1000 Genomes data, across a set of representative SNPs with different VAF s and a range of assumed OR s, we calculated the probability of detection of each study design first, and define those with a probability of detection larger than 0.9 as valid designs. In all the valid designs, we found that the predicted power is in good concordance with the empirical power, although there is a general trend for the prediction model to be conservative (shown in Figure 5).

As an example of finding the optimal design, suppose we have the following values for our cost function: $C_E = \$18,915$, $C_P = \$970$, and $C_X = \$300$ and a total budget of $C_T = \$452,915$ (Dave Duggan from TGen, personal communication). We would like to know the optimal pooling study design for a maximum available sample of 5,000 and an error rate $\varepsilon = 0.01$. The results are shown in Figure 6 for $VAF = 0.03$ and $OR = 2.0$. In the contour plot, we see that the power increases with number of pools and number of individuals per pool, under the constraints of the given costs and sample size. The overlap of the solidlined region (valid study designs) and dashed-lined region (study designs satisfying the cost constraints) define the region of interest. In this example, the optimal power is 0.42, achieved with 51 pools with 62 individuals within each pool, and an average coverage of 25. The top 10 choices of optimal designs are shown in Supporting Information Table SI.

Of course, a pivotal question facing researchers is when to use a pooling study design and when to use individuallevel sequencing. To compare these study designs, we use a similar cost function, substituting the number of individuals for the number of pools and assuming that $4\times$ coverage is sufficient to generate genotyping calls with no errors. With a total available budget of \$452,915, we can individually sequence 200 individuals. For an SNP with a $VAF = 0.03$, and an $OR = 2.0$, the expected power is 0.40 (calculated using Quanto [Gauderman, 2002]).

More generally, to compare a pooling design vs. individual-level sequencing, we investigated power over a range of different cost functions. In such a comparison, it is more informative to focus on the relative value of sample preparation cost vs. sequencing cost (C_P/C_X), since the experiment-wise cost is fixed (C_E). As before, we focus on the scenario with a $VAF = 0.03$, an $OR = 2.0$, and an error rate $\varepsilon = 0.01$. For the cost function, we held constant $C_E = \$10,000$ and $C_P = \$1,000$ and varied the value of C_X from \$50 to \$500. For each value of C_X , we calculated the corresponding cost resulting from an individual-level sequencing design (assuming 200 individuals with $4\times$ coverage), and then used this to constrain the search for an optimal pooling study design. The power for an individual design is constant at 0.40. However, the power for a pooling design varies with cost. Figure 7 demonstrates that as C_X drops, there is a gain in power for an optimal pooling design, with all pooled designs having greater power than the individual sequence design.

A more general set of comparisons is made over a grid of potential $VAFs$ and ORs , with varying values of sequencing cost C_X . As shown in Figure 8, pooling designs provide improved power over individual designs in most of the scenarios, except when VAF is very small ($VAF = 0.005$). However, our assumption that there was no effect of sequencing errors in individual designs might not be true when VAF is as small as 0.005.

We also examined this comparison in a GWAS setting, where the significance level is set to be 5×10^{-8} . In such a setting, pooling was superior to individual sequencing in all the scenarios we investigated, especially when the sequencing cost is low relative to other costs. For example, assuming sequencing cost (C_X) decreases to \$50 while the other costs are constant ($C_E = \$10,000$ and $C_P = \$1,000$), for a variant with an odds ratio of 1.2 and allele frequency 0.2, the power from individual design is close to 0, whereas the optimal pooling design yields power of 0.15; for a variant with an odds ratio of 1.6 and allele frequency 0.1, the power from individual design is 0.32, whereas the optimal pooling design yields power of 0.9.

DISCUSSION

We have shown that pooling designs for genetic association studies using NGS can be a viable alternative to designs with individual-level sequencing. However, careful consideration of the design parameters is needed to understand the tradeoffs between the number of pools, the number of individuals within each pool, and the total coverage per pool. These tradeoffs must also be understood in the context of the goals of the study (e.g., discovery vs. association testing; rare variants vs. common variants) and limitations of budget and sample size.

For testing association, an increase in the number of pools for case-control comparisons can increase power. In addition, increasing the number of individuals within each pool can result in a corresponding increase in power due to a decrease in the variance for the estimated allele frequency for each pool. However, our simulation study demonstrates that the maximum power is not simply a matter of increasing both of these factors until the budget is exhausted and all individuals are included in pools. While increasing the number of individuals within a pool increases the number of potential chromosomes to be sequenced, to ensure that there is a high probability that all chromosomes within a pool are read there must be sufficient total coverage. Moreover, power for testing association with rare variants will be impacted more substantially with low average coverage per chromosome. Thus, under a limited budget and in certain circumstances, there are advantages to decreasing the number of pools while increasing the coverage for each pool. Likewise, increasing the number of individuals within a pool may be counterproductive if limits in coverage have already been reached. In the extreme, this interplay not only impacts power, but also the validity for inference. Too few pools, too few individuals, or inadequate coverage can result in unstable estimates and invalid tests of association.

Future studies need more than broad conclusions for guidance. To accomplish this, we developed a procedure for determining the optimal study design given total available budget and sample size, as well as the sequence error rate, VAF , and odds ratio for the targeted

causal SNP. In searching for the optimal study design, there are two key steps in the procedure. First, a proposed study design is determined to be valid if the design has a probability of detection greater than 0.9. As with power, too few pools, too few individuals, or inadequate coverage results in lower probabilities of detection. However, since the impact of these design parameters is complicated, the ability to calculate probabilities of detection can be useful if the goal of a particular study is discovery and not association. As highlighted in the Appendix and corresponding figures (Supporting Information Figures S1, S2, and Table SII), pooling designs outperform individual-level sequencing for discovery of rare variants even when accounting for sequencing error. The second step in our power estimation procedure uses a regression model for predicting the expected noncentrality parameters and, in turn, the expected power. The results from over 116,000 simulation scenarios were used to construct the prediction model. Overall, the prediction model does very well in predicting the empirical power for the range of scenarios given by the simulations. This approach was used since an exact calculation of the noncentrality parameter is infeasible due to the complexity of our multilevel model and the Bayesian framework in which our estimation is done, and the prediction model allows for flexibility in predicting power for novel study designs. However, there will certainly be an increase in the uncertainty when predicting power for study designs not within the simulated ranges of design parameters. For example, prediction of a pooled GWAS analyses using a α -level of 5×10^{-8} will have increased uncertainty since very few simulated designs resulted in test statistics that reach such significance.

The ability to predict power for a novel pooled study design facilitates comparison to individual-level sequencing designs. Determination of which design is most powerful relies heavily on the cost function. Here, we used a simple cost function that is driven by a sample cost (e.g., cost per pool [or individual]) and a sequencing cost (e.g., cost per unit coverage per pool [or individual]). Under this cost function, the relative value of sample cost vs. sequencing cost (C_P/C_X) is most informative for determining which design option is best. Our results indicate that even as sequencing costs drop, pooled designs offer potentially more power in many situations in comparison to cost-equivalent individual designs. This is mainly driven by the sample costs and highlights how the goals of the study can impact design choice. For targeted sequencing studies, sample costs are often high relative to sequencing costs due to library formation. In such studies, pooling may be the best design. In contrast, whole-genome sequencing studies currently have relatively low sample costs with much higher sequencing costs. For these studies, individual-level designs may be more appropriate, especially for rare variants. To present comparisons that would highlight trends, we assumed that 4× coverage was sufficient to obtain genotypes without error from individual-level sequencing. We ignored sequencing error and variation in coverage along the genome, as well as techniques that can improve genotype calling from sequencing data such as imputation. Despite these assumptions, we believe that our approach provides a fair comparison between the design options and clearly presents the influencing factors.

Finally, for testing the marginal association of a variant, our novel hierarchical modeling approach accounts for the number of pools, the number of individuals within each pool, the coverage, and the sequencing error. We have presented the model to test association between cases and controls, but the general framework is applicable to the comparison of

any categorical variable that can be conditioned on to construct distinct pools, such as tumor vs. normal. Additionally, one can construct pools by conditioning on more than one variable and then obtain marginal estimates of association for any single variable by simply including them as covariates in the association model at the pool-level. Such an approach may be sufficient for the adjustment of potential confounding due to population stratification by constructing case and control pools where each pool consists of individuals with similar values for key ancestry components. Sequencing studies have the potential for substantial impact in the discovery and analysis of rare variants. Due to sparse data, most analysis approaches for rare variants have relied on a burden test where the number of rare variants for each individual is summed and a resulting overall risk index is tested for association [Madsen and Browning, 2009]. In its current form, our approach focuses only on testing the marginal association for each variant. However, one could potentially extend our approach to include a test of the difference between the scaled sum of the variant frequencies in case pools vs. control pools. Further research is certainly warranted in this area.

For many studies, there simply is no substitute for individual-level sequencing as it offers flexibility to test multiple hypotheses (e.g., tests of interaction) or allows for association testing when individuals are correlated such as in family studies. As sequencing costs continue to drop, the ability to obtain individual-level data may be too enticing to warrant pooled approaches. However, if the goals are discovery or association testing, pooled designs will continue to be a viable option. In addition, there is a rich literature on various two-stage options in which one can potentially use pooling as the first stage to accomplish both discovery of novel variants and ranking of those variants via association tests for second-stage follow-up via genotyping [Zhao and Wang, 2009]. New technologies may also facilitate more advanced designs such as pools of individuals within barcoded pools all sequenced in a single run [Smith et al., 2010].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by grants R01CA140561 and P50HG002790 from National Institutes of Health (NIH), R01ES019876 from NIH/National Institute of Environmental Health Sciences (NIEHS), P30ES007048 from NIEHS, R01HL087680 from the National Heart, Lung, and Blood Institute (NHLBI), R01CA143237 from NIH/National Cancer Institute (NCI), and R01MH084678 from the National Institute of Mental Health (NIMH). The authors would also like to thank Christopher Edlund for his advice on processing 1000 Genomes data.

APPENDIX

VARIANT DETECTION: COMPARISON BETWEEN POOLING VS. INDIVIDUAL DESIGNS

We also investigated the use of a pooling design for the discovery of novel variants, based on the probability of detection proposed in the Methods section. As an example, with the same cost function, suppose we were interested in a variant with $VAF = 0.01$, in an individual sequencing design we can sequence 200 individuals with $4\times$ coverage (again

assuming that is enough to yield perfect sequencing result), the probability of discovery would be $1 - (1 - 0.01)^{200} = 0.866$. With the same sequencing cost (and with no more than 4,000 individuals), the optimal pooling design would yield a probability of detection is 0.948. The discovery probability of different study designs was shown in Supporting Information Figure S1.

Amore general comparison between pooling vs. individual sequencing is made with varying *VAFs* and cost functions, with the same cost for 4× coverage on 100 individuals. In Supporting Information Figure S2, we can see that even when *VAF* is very small, optimal pooling design yields a very high detection probability. The optimal designs are summarized in Supporting Information Table SII.

References

- Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, Fay JC, Mitra RD. Quantification of rare allelic variants from pooled genomic DNA. *Nat Meth.* 2009; 6:263–265.
- Erich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, Hannon GJ. DNA Sudoku—harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Res.* 2009; 19:1243–1253. [PubMed: 19447965]
- Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet.* 2009; 10:241–251. [PubMed: 19293820]
- Futschik A, Schlotterer C. Massively parallel sequencing of pooled DNA samples—the next generation of molecular markers. *Genetics.* 2010; 186(1):207–218. [PubMed: 20457880]
- Gauderman W. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol.* 2002; 25(4):327–338.
- Kim SY, Li Y, Guo Y, Li R, Holmkvist J, Hansen T, Pedersen O, Wang J, Nielsen R. Design of association studies with pooled or unpooled next-generation sequencing data. *Genet Epidemiol.* 2010; 34:479–491. [PubMed: 20552648]
- Lee JS, Choi M, Yan X, Lifton RP, Zhao H. On optimal pooling designs to identify rare variants through massive resequencing. *Genet Epidemiol.* 2011; 35:139–147. [PubMed: 21254222]
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83:311–321. [PubMed: 18691683]
- Lupton MK, Proitsi P, Danillidou M, Tsolaki M, Hamilton G, Wroe R, Pritchard M, Lord K, Martin BM, Kloszewska I, Soininen H, Mecocci P, Vellas B, Harold D, Hollingworth P, Lovestone S, Powell JF. Deep sequencing of the nicastrin gene in pooled DNA, the identification of genetic variants that affect risk of Alzheimer's disease. *PLoS ONE.* 2011; 6:e17298. [PubMed: 21364883]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol.* 2009; 34:188–193. [PubMed: 19810025]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet.* 2011; 7:e1001322. [PubMed: 21408211]
- Out AA, van Minderhout IJHM, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, Weigel D, van Galen M, Taschner PEM, Tops CMJ, Breuning MH, van Ommen G-JB, Dunnen den JT, Devilee P, Hes FJ. Deep sequencing to reveal new variants in pooled DNA samples. *Hum Mutat.* 2009; 30:1703–1712. [PubMed: 19842214]
- Prabhu S, Pe'er I. Overlapping pools for high-throughput targeted resequencing. *Genome Res.* 2009; 19:1254–1261. [PubMed: 19447964]

- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86:832–838. [PubMed: 20471002]
- Quintana MA, Berstein JL, Thomas DC, Conti DV. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. *Genet Epidemiol.* 2011; 35:638–649. [PubMed: 22009789]
- Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008; 26:1135–1145. [PubMed: 18846087]
- Shental N, Amir A, Zuk O. Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Research.* 2010; 38(19):e179. [PubMed: 20699269]
- Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, Nislow C. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* 2010; 38:e142–e142. [PubMed: 20460461]
- Wang T, Pradhan K, Ye K, Wong L-J, Rohan TE. Estimating allele frequency from next-generation sequencing of pooled mitochondrial DNA samples. *Front Genet.* 2011; 2:51. [PubMed: 22303347]
- Zhao Y, Wang S. Optimal DNA pooling-based two-stage designs in case-control association studies. *Hum Hered.* 2009; 67:46–56. [PubMed: 18931509]

Fig. 1.

DAG of the overall hierarchical model, focusing on a single marker in a single pool. In the unobserved pool-level, given the total number of individuals (N_p) and the allele frequency (ρ_p), the number of variant-carrying chromosomes (Z_p) is distributed as Binomial (N_p, ρ_p). The allele frequency (ρ_p) is connected with case-control status (W_p) in a logistic model. Given the coverage depth (X_p), the number of variant reads (V_p) is distributed as Binomial (X_p, θ_p), where θ_p is a mixed effect of sequencing error (ϵ) and the proportion of variant-carrying chromosomes in the pool ($Z_p/(2N_p)$). To mimic region-capturing variation, the coverage depth (X_p) is simulated to be Poisson (λ, γ_p), where γ_p is distributed as Gamma ($\alpha, 1/\alpha$). It is not applied in the analysis model, where X_p is observed data.

A The Validity of a Study Design with $\lambda=40$ and $VAF=0.05$ B The Validity of a Study Design with $\lambda=160$ and $VAF=0.05$ **Fig. 2.**

Summary of the validity of study designs of a variant with $VAF=0.05$. A study design is invalid if the type I error is not uniformly distribution (based on K-S test). In plot A, with relatively low coverage ($\lambda = 40$), many designs are invalid, especially those with small number of pools (P) and larger number of individuals per pool (N_p). In plot B, due to a higher coverage ($\lambda = 160$), most of the designs are valid.

Fig. 3. Comparison of empirical power in different scenarios and designs. Each plot shows the power (y-axis) of finding the association with corresponding coverage (λ , in different line types) and number of individuals per pool (N_p , in x-axis). In plots A and B, we focus on a variant with $VAF = 0.05$, $OR = 1.6$ and assumed sequencing error $\epsilon = 0.01$, whereas in plots C, D, and E, a more rare variant with $VAF = 0.01$ and $OR = 2$ are investigated. We include both valid and invalid designs to demonstrate the overall trend of power.

Fig. 4.

General trends of empirical power with respect to the scenario and design parameters. Each plot shows the power (contour gray-scale), fixing one of the three design parameters (P , N_p , or λ) while changing the other two. A variant with $VAF=0.05$ and $OR=1.6$ is investigated in plots A, B, and C, while a more rare variant with $VAF=0.01$ and $OR=2$ is investigated in plots D, E, and F.

Fig. 5.

Prediction of power in semisimulation studies using 1000 Genomes data. Using the 1000 Genomes exome alignment data, focused on a set of representative SNPs, assumed Ors, and study designs, the predicted power is calculated directly using the prediction model, whereas the empirical power is calculated based on semisimulation studies.

Fig. 6.

Grid search for the optimal pooling design, for a study interested in a variant with $VAF=0.03$ and $OR=2$, assuming a fixed cost function, with constraints on cost and sample size. The contour plot shows the predicted power (with maximal coverage possible) over a range of number of pools (y -axis) and number of individuals per pool (x -axis). The region outside of the dashed lines are designs that are impossible to achieve due to cost constraint, and the region outside of the solid lines are invalid designs or designs that exceed the maximum sample size.

Fig. 7.

The power of the top 10 pooling designs (y -axis) for a specific study ($VAF = 0.03$, $OR = 2$), with the same sample size constraint, different cost functions (x -axis) but comparable cost constraints with individual designs. Assuming that in the simplified cost function $C_T = C_E + C_P P + C_X P \lambda$, we fix the experiment-wise cost C_E (at \$10,000), and the sample cost C_P (at \$1,000), with varying sequencing cost C_X . We find total cost C_T for sequencing 200 individuals at $4\times$, and use the corresponding C_T as cost constraint to find the optimal pooling design. In all the plots, the power for individual designs is fixed at 0.4 (shown as the dashed line), whereas the optimal power for pooling designs vary with different cost functions.

Fig. 8.

Power (y-axis) comparison of optimal pooling designs with various cost function (symbol types), vs. individual designs (dashed line), over a range of variant allele frequencies (x-axis). The odds ratio is assumed to be 1.6 and the cost function is the same as in Figure 7.

TABLE I
The SNPs in the semisimulation study using 1000 Genomes data

Chromosome	Position	VAF in sample	RS no.
2	272051	0.03	rs11553742
2	677295	0.08	rs3210390
2	1204796	0.16	rs6751090

TABLE II
Summary of the partial R^2 of the prediction models. Four different models are built for different variant allele frequency range

Design parameters	Model 1: $VAF = [0.002, 0.01]$	Model 2: $VAF = [0.01, 0.05]$	Model 3: $VAF = [0.05, 0.1]$	Model 4: $VAF = [0.1, 0.2]$
VAF	0.1142	0.0150	0.0011	0.0240
$\ln(OR)$	0.4069	0.3941	0.3474	0.3132
ε	0.0062	0.0024	0.0066	0.0004
P	0.0724	0.1486	0.2206	0.2683
λ	0.1607	0.1426	0.1419	0.1622
N_p	0.07	0.1294	0.0890	0.0761
$P^{1/3}$	0.0095	0.0225	0.0377	0.0471
$\lambda^{1/3}$	0.0486	0.0384	0.0561	0.0613
$N_p^{1/3}$	0.0335	0.0500	0.0391	0.0338
$[\ln(OR)]^2$	0.0082	0.0100	0.0100	0.0093
$VAF * \ln(OR)$	0.0018	0.0001	0.0000	0.0011
$\lambda * N_p$	0.0121	0.0123	0.0121	0.0108
$VAF * P$	0.0007	0.0002	0.0000	0.0003
$VAF * \lambda$	0.0011	0.0000	0.0001	0.0000
$VAF * N_p$	0.0031	0.0001	0.0002	0.0004
$\ln(OR) * P$	0.0015	0.0022	0.0030	0.0027
$\ln(OR) * \lambda$	0.0029	0.0030	0.0030	0.0014
$\ln(OR) * N_p$	0.0003	0.0003	0.0006	0.0000
$\varepsilon * P$	0.0000	0.0000	0.0002	0.0000
$\varepsilon * \lambda$	0.0001	0.0000	0.0000	0.0000
$\varepsilon * N_p$	0.0012	0.0006	0.0001	0.0000
$VAF * \lambda * N_p$	0.0000	0.0000	0.0000	0.0000
$\ln(OR) * \lambda * N_p$	0.0001	0.0000	0.0001	0.0001
$VAF * \ln(OR) * P$	0.0000	0.0000	0.0000	0.0000
$VAF * \ln(OR) * \lambda$	0.0000	0.0000	0.0000	0.0000
$VAF * \ln(OR) * N_p$	0.0000	0.0001	0.0000	0.0000
$\varepsilon * \lambda * N_p$	0.0000	0.0000	0.0001	0.0000
$VAF * \ln(OR) * \lambda * N_p$	0.0000	0.0000	0.0000	0.0000