BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 14 2011, pages 1995–1997
doi:10.1093/bioinformatics/btr305

Genome analysis

Advance Access publication June 2, 2011

Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association

Yufeng Shen[1,2,*], Ruijie Song[1] and Itsik Pe'er[1,2,*]

[1]Department of Computer Science, Columbia University, New York, NY 10027 and [2]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Whole-genome sequencing (WGS) allows direct interrogation of previously undetected uncommon or rare variants, which potentially contribute to the missing heritability of human disease. However, cost of sequencing large numbers of samples limits its application in case–control association studies. Here, we describe theoretical and empirical design considerations for such sequencing studies, aimed at maximizing the power of detecting association under the constraint of study-wide cost.

Results: We consider two cost regimes. First, assuming cost is proportional to the total amount of base pairs to be sequenced across all samples, which is a practical model for whole-genome sequencing, we explored the tradeoff in terms of study power between increasing the number of subjects and increasing depth coverage. We demonstrate that the optimal power of detecting association is achieved at medium depth coverage under a wide range of realistic conditions for case-only sequencing designs. Second, if cost is fixed per sample, which is approximately the case in exome sequencing, we show that in a simple case+control sequencing study, the optimal design should include cases totaling $1/e$ of all subjects.

Availability: A web tool implementing the methods is available at http://www.cs.columbia.edu/~itsik/OPERA/.

Contact: yshen@c2b2.columbia.edu; itsik@cs.columbia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 7, 2010; revised on May 1, 2011; accepted on May 11, 2011

1 INTRODUCTION

Genome-wide association studies (GWAS) have greatly improved our understanding of the genetics of human diseases (Altshuler *et al*., 2008). However, for the majority of common diseases, substantial fractions of heritability remain to be explained. Rare variants, which are under the radar of GWAS, are suggested to have significant contributions to missing heritability (Eichler *et al*., 2010; Frazer *et al*., 2009; Yang *et al*., 2010).

Complete ascertainment of rare variants requires genome-wide resequencing, which is now feasible thanks to the development of next-generation sequencing technologies (Bentley *et al*., 2008). Exome sequencing (Ng *et al*., 2009) and whole-genome sequencing (WGS) (Lupski *et al*., 2010) have been successfully applied to identify rare causal variants of Mendelian diseases. In these studies, a causal gene was implicated if it harbors rare functional mutations that are shared among cases but absent in controls or single nucleotide polymorphism (SNP) databases. Extending this approach to association studies of complex diseases requires more than a handful of samples, because the causal variants of such conditions only have statistical effect (Cohen *et al*., 2006). Therefore, it is necessary to carefully consider the design of such studies under the constraint of limited resources.

A critical technical attribute of WGS data is the number of times each site is observed in a sequencing read, usually termed the depth-of-coverage. Depth-of-coverage is a key determinant of quality for sequencing information, and in particular, w.r.t. rare variants. Accuracy and completeness of detection and calling such variants from sequencing depend on high depth-of-coverage, due to the randomness of read placement (Lander and Waterman, 1998) and non-negligible error rate (Mckernan *et al*., 2009). Higher depth coverage for variant discovery and genotyping is attainable by using more sequencing resources (e.g. Illumina lanes) per sequenced sample. However, given a fixed budget, more sequencing per sample is in conflict with analyzing a large sample size, which is required for adequate statistical power of detecting associations. Here, we propose to resolve this tradeoff by finding the design that maximizes the power for detecting associations of rare variants. We define rare variants as the ones with minor allele frequency (MAF) $< 1\%$ in the population, and uncommon variants as the ones with MAF between 1 and 5%. Here, we use the term rare variants to include both rare and uncommon variants.

2 METHODS AND RESULTS

2.1 Sequencing cases with cost proportional to total bases

We first assume a cost regime where the overhead cost of a subject is negligible comparing to that of the sequenced bases. Under fixed total cost and therefore affordable total bases (T) is fixed, genome-wide average depth coverage λ is determined by the number of cases to be sequenced (N_1): $\lambda = T/N_1$. For a rare variant, we assume homozygous carriers are rare enough to be ignored. The power of correctly calling a heterozygous carrier from sequencing data is a function of λ: $R(\lambda)$, usually with an approximate sigmoid. It can be determined from empirical data (Supplementary Fig. S1) or approximated analytically (Wendl and Wilson, 2009; Wheeler *et al*., 2008).

Allelic heterogeneity is likely to be ubiquitous among genes that harbor causal rare variants (Bodmer and Bonilla, 2008; Pritchard, 2001). Grouping rare variants according to genes or pathways they disrupt for association testing (Price *et al*., 2010) is necessary to increase power. For simplicity, we

*To whom correspondence should be addressed.

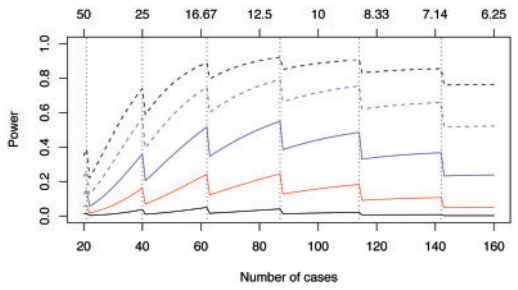

Fig. 1. Power versus the number of cases (N_1) under budget constraint, assuming $T=1000\times$, $N_0=400$, $K_0=0$ and $Q=5\times10^{-6}$. Different horizontal lines represent power assuming different case carrier frequencies (F_1), with solid black, solid red, solid blue, dashed blue and dashed black representing $F_1=0.05, 0.075, 0.1, 0.125$ and 0.15, respectively. The unit on the top x-axis is average depth coverage corresponding to the number of cases.

consider a gene as disrupted in a sample if it harbors any of the causal variants, and use the collapsing method (Li and Leal, 2008) to test the association.

In the presence of allelic heterogeneity, we assume there are M rare causal variants within a gene (or pathway). These alleles are independently distributed in the population, with respective carrier frequencies $h_1,\ldots,h_M$ among cases. The carrier status for any of these rare variants is a Bernoulli variable with compound carrier frequency $p=R(\lambda)\left(1-\prod_{i=1}^{m}\left(1-h_i\right)\right)$ in cases. Let $F_1\equiv\sum_{i=1}^{M}h_i$, then $p\approx F_1R(\lambda)$ when $h_i<<1$, and the number of observed carriers K_1 among N_1 cases follows a binomial distribution with parameters $(N_1, F_1R(\lambda))$. An economic design is to sequence cases only and utilize publicly available data for additional N_0 samples as controls. For example, we use $N_0=400$, which represents a lower bound of the sample size from one major population of the 1000 Genomes Project (1000 Genomes Project Consortium *et al*., 2010). Given T, N_0, F_1 and $R(\lambda)$, we provide an online tool (OPERA) that can calculate the power of detecting association from different values of N_1 (and λ) under certain Type I error cutoff (Q). As an example, we show the results with resources of $T=1000\times$ genomes under a simplistic assumption that the number of carriers in controls (K_0) is 0 (Fig. 1). The power curve takes the unusual shape of a sawtooth wave that reflects discrete artifacts (Supplementary Material) of the test statistics for association of rare variants (Supplementary Figs S2 and S3), superimposed on a smoothed concave curve through the respective tips of the sawteeth. Moving along the x-axis, the smoothed curve initially increases with N_1 when N_1 is relatively small, reflecting the fact that when N_1 is small, λ is large, therefore $R(\lambda)$ is close to 1 and changes very little as N_1 increases. For larger sample sizes and lower coverage, where $R(\lambda)$ is away from 1, increasing N_1 further reduces λ and $R(\lambda)$ sufficiently to negate the benefit of increasing N_1, thus power starts to decrease from one sawtooth tip to the next. We present similar results with $K_0=1$ (Supplementary Fig. S4a), which can represent singleton variants in controls sequenced at low coverage. In general, for a gene with compound carrier frequency F_0 of rare functional variants in controls, the expected power is the sum of power under possible K_0 values weighted by the probability of observing K_0. We show the power curve with hypothetical $F_0=0.001$ (Supplementary Fig. S4b).

2.2 Fixed sequencing cost per case or control sample, sequencing both cases and controls

The other cost regime is where per-sample overhead is large enough to be approximated as fixed cost regardless of coverage, which is approximately the case for practical exome sequencing. The budget then translates to a total number of subjects to be sequenced. The remaining decision for the investigator is what fraction of these subjects should be cases versus controls. While variants may be partially penetrant, either causal or protective (Neale *et al*., 2011; Yi and Zhi, 2011), a simplifying assumption of gene disruptions to be fully penetrant and causal is helpful both in practical exome sequencing (Ng *et al.*, 2009) as well as for theoretical analysis: we prove (Supplementary Material) that in such scenarios the optimal ratio of cases to all samples is $1:e$ (e is the base of the natural logarithm). If disruptions are to be sought in both cases and control without bias, this asymmetry breaks, and naturally one needs to sequence a similar number of subjects from either group.

3 CONCLUSIONS AND DISCUSSIONS

We discussed optimal designs of association studies focused on rare variants using high-throughput resequencing under budget constraint. Under a scenario where the cost is proportional to the total number of sequenced bases and only sequencing cases, there are two important characteristics of the optimal design under a constant budget. First, the curve of power versus the number of cases is not smooth, but resembles a sawtooth. Second and more importantly, smoothing the sawtooth tips, the maximum power of detecting associations is achieved at a medium coverage (λ) where the power of calling heterozygous variants $R(\lambda)$ from sequencing data is suboptimal. Under a second scenario where the cost is proportional to the number of subjects, when assuming sequencing both cases and controls for detecting case-only disruptions, the optimal fraction of cases among sequenced subjects is $1/e$.

Our analysis of the optimal design assumes using simple collapsing method to test the association of rare variants in the presence of allelic heterogeneity. Although some other methods are better powered under certain conditions (Madsen *et al*., 2009; Neale *et al*., 2011; Price *et al*., 2010; Yi and Zhi, 2011), the collapsing method is often the first statistical test researchers conduct after getting new data. Moreover, it has been showed to be empirically adequate in practice when combined with appropriate functional assessment of the variants (Cohen *et al*., 2006; Surolia *et al*., 2010). Our power calculator assumes this test following to traditional calculations that typically focus on a basic scenario and test, rather than a full probabilistic model.

We have implemented a web tool for computing the power under a flexible set of assumptions considering the general cost regime, where both a fixed, per-sample cost and a variable, per coverage price are accrued. This tool could provide investigators means to plan experiments in the practical regimes of budgets and technologies they are facing within their institutions.

Funding: National Science Foundation CAREER 0845677 (to I.P.); National Institute of Health U54CA121852 (to I.P.); International Serious Adverse Event Consortium (to Y.S.).

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al.* (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

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

Cohen,J.C. *et al.* (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.*, **354**, 1264–1272.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Frazer,K.A. *et al.* (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–251.

Lander,E.S. and Waterman,M. (1998) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Lupski,J.R. *et al.* (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.*, **362**, 1181–1191.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Mckernan,K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.

Neale,B.M. *et al.* (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.

Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature,* **461**, 272–276.

Price,A.L. *et al.* (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Pritchard,J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.

Surolia,I. *et al.* (2010) Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. *Nature*, **466**, 243–247.

Wendl,M.C. and Wilson,R.K. (2009) The theory of discovering rare variants via DNA sequencing. *BMC Genomics*, **10**, 485.

Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

Yi,N. and Zhi,D. (2011) Bayesian analysis of rare variants in genetic association studies. *Genet. Epidemiol.*, **35**, 57–69.

1000 Genomes Project Consortium *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.