Journal of Human Genetics (2012) 57, 676–678
© 2012 The Japan Society of Human Genetics All rights reserved 1434-5161/12 $32.00
www.nature.com/jhg

SHORT COMMUNICATION

NDesign: software for study design for the detection of rare variants from next-generation sequencing data

Yuki Sugaya[1], Yasuaki Akazawa[2], Akira Saito[1] and Shigeo Kamitsuji[1]

We developed a software program, NDesign, for the design of a study intended for detecting rare variants from next-generation sequencing (NGS) data. In this study design, the optimal depth of coverage and the average depth of coverage are first evaluated, and then the ability of the designed experiment to obtain a desired power is determined. NDesign has been developed to calculate both these depths, as well as to evaluate the power of the designed experiment. It has a simple implementation in the JavaScript language, and is expected to enable researchers to design optimal NGS studies.

Journal of Human Genetics (2012) **57,** 676–678; doi:10.1038/jhg.2012.81; published online 12 July 2012

Keywords: next-generation sequencing; rare variant detection; study design

INTRODUCTION

Genome-wide association studies (GWAS) have revealed numerous associations between diseases and alleles of single-nucleotide polymorphisms (SNPs).[1–4] While a comprehensive study of the genome can detect an association with high sensitivity, studies are typically limited to finding SNPs with common or moderate frequency in a given population. Although variants with low frequency might also be responsible for disease, detection of the variants is not realistic by GWAS, because it is hard to obtain sufficient sample size for a desired power, and it is hard to distinguish such results from experimental errors. Recent advances in high-throughput sequencing technologies, known as next-generation sequencing (NGS) technologies, can potentially identify such associations, using a parallel short-read strategy for DNA sequencing. A number of short reads are aligned over each locus; therefore, the observed alleles are useful in distinguishing the results from experimental errors. If an abnormal allele is observed only in specific disease patients, it is considered to be associated with the disease. This DNA sequencing approach can detect variant loci having a low frequency of occurrence, as well as those with a moderate or high frequency. As many common variants have already been detected by GWAS, we focus on the detection of rare variants in this study. With the emergence of commercially available platforms, several associations have already been identified by NGS;[5–7] however, an issue remains in that the choice of depth of coverage in the design of a study has not yet been discussed in adequate detail.

A well-considered study design is necessary for conducting an experiment successfully and economically. A number of reads are required to be aligned at a locus for determining whether it is a variant locus because of possible calling errors. To obtain sufficient observation of reads at a locus, known as depth of coverage, to identify the variant locus, a greater number of total sequences are necessary, which would unfortunately increase cost. One of the ways of obtaining results feasibly is to design the minimum indispensable depth to identify the variant. An approach to study design has already been proposed[8] for this purpose; however, this approach is complicated in that it introduces a negative binomial distribution for the depth of coverage, and calculates power via a simulation. Further, the method is not readily available, as it has not been implemented in any software program. We herein introduce a simpler model for power and a software implementation of the study design method. The power to detect variants can be explicitly formalized in terms of the significance level, the calling error probability and the probability of observing variant alleles based on the binomial distribution; consequently, the proposed study design can be considered as the design of an experiment in which the average depth of coverage exceeds the optimal depth of coverage derived from the calculated power. We have developed a software program termed NDesign to calculate the optimal and average depths of coverage, and to evaluate the feasibility of the designed experiment. NDesign has a simple implementation in JavaScript, and we believe that it will benefit researchers attempting to detect rare variants from the NGS data.

METHODS

Design of optimal depth of coverage

Rare variant detection within an individual. First, we derive an explicit formula for the power to detect a rare variant at a locus within an individual. Here, we assume that the carrier of the variant allele is a heterozygote of the variant and normal alleles, because the frequency of the variant is assumed to

[1]Statistical Genetics Analysis Division, StaGen Co., Ltd, Tokyo, Japan and [2]Department of Electrical Engineering and Bioscience, School of Advanced Science and Engineering, Waseda University, Tokyo, Japan
Correspondence: Y Sugaya, Statistical Genetics Analysis Division, StaGen Co., Ltd, KUGA Building 8F, 4-11-6, Kuramae, Taito-ku, Tokyo 1110051, Japan.
E-mail: sugaya@stagen.co.jp

Received 24 February 2012; revised 7 June 2012; accepted 8 June 2012; published online 12 July 2012

be low. When we observe D alleles at a locus for the carrier, the number of observations of the rare variant follows a binomial distribution $B(D, p)$, where p is the probability of observing the variant, which can be taken to be equal to 1/2 in the case of rare variant detection within an individual. For a non-carrier, the number of observations follows a different binomial distribution $B(D, p_{\mathrm{error}})$, where p_{error} is the calling error probability of observing the variant allele from the homozygote individual of the normal allele, which normally takes a value lower than p. Upon setting the significance level as α, the power to detect the variant can be described as

$$\text{Power} = \sum_{x=x*(\alpha)}^{D} B(x; D, p), \qquad (1)$$

where $x^*(\alpha)$ is the critical number of rare variant observations, and

$$x*(\alpha) = \min\{x \mid \sum_{i=x}^{D} B(i; D, p_{\mathrm{error}}) \leqslant \alpha\}.$$

Our first goal is to determine the optimal depth of coverage, d_{optim}, as the minimum depth exceeding the desired power derived from equation (1). This depth is indispensable in identifying the variant allele with this desired power.

Rare variant detection within pooled sample. The extension of this discussion to pooled sample data is simple. If we assume that n carriers of the variant allele exist in the pool, we simply replace the probability of observing rare variants, $p = 1/2$, with $p = n/2N$, where N is the pool size. The optimal depth of coverage for pooled sample data at a particular desired power can also be evaluated.

Design of experiment

Our second goal is to calculate the average depth of coverage for the designed experiment, after which the experiment can be evaluated by examining whether the average depth exceeds the obtained optimal depth of coverage. For simplicity, we assume that all reads are uniformly aligned over the genes or regions targeted in the study. Therefore, the average depth of coverage can be explicitly expressed as

$$d_{\mathrm{avg}} = \frac{L}{lN},$$

where L is the total sequence for the employed sequencer and sequencing method, and l is the length of the target genes. In the Discussion section, we discuss the case in which we assume a non-uniform alignment of the reads. The total sequence L can be expressed as $L = br$, where b is the number of beads (or clusters) per experiment (one run) and r is the read length. The parameters for well-known commercially available NGS platforms are summarized in Table 1. The total sequences are also listed in this table. The feasibility of conducting the designed experiment can be evaluated by comparing the obtained average depth of coverage with the optimal depth of coverage.

Table 1 Summary of read length, number of beads and total sequence for four well-known commercially available NGS platform

	Hiseq2000	*GAIIx*	*5500 SOLiD*	*GS FLX*
Read length				
Single-end read	100	35	75	400
Paired-end read	100 × 2	75 × 2	75 + 35	400
No. of beads (or clusters) per run	3.00E + 09	3.20E + 08	7.50E + 08	1.00E + 06
Total sequence				
Single-end read	3.00E + 11	1.12E + 10	5.63E + 10	4.00E + 08
Paired-end read	6.00E + 11	4.80E + 10	8.25E + 10	4.00E + 08

Total sequence is calculated from read length and number of beads.

AVAILABILITY AND IMPLEMENTATION

We have developed a software program, NDesign, to determine the optimal depth of coverage with desired power and the average depth of coverage for the designed experiment. NDesign and its user guide are available free of charge at http://www.stagen.co.jp/ndesign.html. This program is written in JavaScript and can therefore run on several standard Web browsers that can interpret this language.

DISCUSSION

We have proposed a binomial-distribution-based study design method for the detection of rare variants from NGS data. A good approximation for the power may be provided by using another probability distribution; however, we obtained an exact formula using the binomial distribution without considering any approximation conditions. The probability of observing rare variants may fluctuate, owing to several well-known biases (for example, duplication bias for read, alignment error or GC contents). However, the expected power can be computed without any such biases, because the employed probability corresponds to the expected one.

We have assumed that the alignment of reads is uniform over the target genes or regions. However, the actual alignment is not uniform; that is, the depth has a distribution over the target genes. An optimal experiment is, of course, one in which the depth of coverage at every locus exceeds the evaluated optimal depth of coverage; in other words, the experiment with $\beta(d) = 1$ is the optimal one, where

$$\beta(d) = \frac{1}{l} \sum_{\mathrm{t;target\ genes}} 1_{\{d(\mathrm{t}) \geqslant d_{\mathrm{optim}}\}},$$

$d(\mathrm{t})$ is the depth of coverage at locus t, and

$$1_{\{d(\mathrm{t}) \geq d_{\mathrm{optim}}\}} = \begin{cases} 1 & \text{if } d(\mathrm{t}) \geqslant d_{\mathrm{optim}} \\ 0 & \text{otherwise} \end{cases}.$$

The summation runs over the target genes of the experiment. The experiment having a uniform distribution, $d(\mathrm{t}) = d_{\mathrm{avg}}$, would therefore be optimal, provided $d_{\mathrm{avg}} > > d_{\mathrm{optim}}$. In the case that is difficult to assume the distribution before the experiment, the criterion $d_{\mathrm{avg}} > d_{\mathrm{optim}}$ may provide a feasible evaluation of the designed experiment. It will be better to use an empirical distribution instead of a distribution based on a probabilistic model if it is available, because there are sequence-specific biases affecting the distribution, which depend on the characteristics of the reagents used and platform-specific chemistry. In the case that an experiment with low $\beta(d)$ has already been conducted, this information can be used as the basis for design of an additional experiment to improve β, using the obtained empirical distribution. The introduction of an empirical distribution into NDesign may be considered in the future work. Currently, we have developed this software by assuming a uniform distribution to realize a simple implementation; this assumption is presently adequate for planning an experiment in the early stages of a study.

ACKNOWLEDGEMENTS

We thank the anonymous referees for their valuable comments and suggestions to improve the quality of the manuscript. We are also grateful to Dr Ian Saunders at the Commonwealth Scientific and Industrial Research Organisation for careful reading our manuscript and helpful comments.

1 Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32,** 650–654 (2002).
2 Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. K., Sackler, R. S., Haynes, C. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308,** 385–389 (2005).
3 Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314,** 1461–1463 (2006).
4 Hampe, J., Franke, A., Rosenstiel, P., Till, A., Teuber, M., Huse, K. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nat. Genet.* **39,** 207–211 (2006).
5 Hoischen, A., van Bon, B. W., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M. *et al.* *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42,** 483–485 (2010).
6 Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L. *et al.* Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* **362,** 1181–1191 (2010).
7 Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M. *et al.* Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).
8 Sampson, J., Jacobs, K., Yeager, M., Chanock, S. & Chatterjee, N. Efficient study design for next generation sequencing. *Genet. Epidemiol.* **35,** 269–277 (2011).