Genetic Epidemiology 35:269–277 (2011)

Efficient Study Design for Next Generation Sequencing

Joshua Sampson,[1]* Kevin Jacobs,[2] Meredith Yeager,[2] Stephen Chanock,[2] and Nilanjan Chatterjee[1]

[1]*Biostatistics Branch, DCEG, National Cancer Institute, Rockville, Maryland*
[2]*Core Genotyping Facility, DCEG, National Cancer Institute, Gaithersburg, Maryland*

Next Generation Sequencing represents a powerful tool for detecting genetic variation associated with human disease. Because of the high cost of this technology, it is critical that we develop efficient study designs that consider the trade-off between the number of subjects (n) and the coverage depth (μ). How we divide our resources between the two can greatly impact study success, particularly in pilot studies. We propose a strategy for selecting the optimal combination of n and μ for studies aimed at detecting rare variants and for studies aimed at detecting associations between rare or uncommon variants and disease. For detecting rare variants, we find the optimal coverage depth to be between 2 and 8 reads when using the likelihood ratio test. For association studies, we find the strategy of sequencing all available subjects to be preferable. In deriving these combinations, we provide a detailed analysis describing the distribution of depth across a genome and the depth needed to identify a minor allele in an individual. The optimal coverage depth depends on the aims of the study, and the chosen depth can have a large impact on study success. *Genet. Epidemiol.* 35:269–277, 2011.

Key words: next generation sequencing; sequencing depth; study design; rare variants

Additional Supporting Information may be found in the online version of the article.
*Correspondence to: Joshua Sampson, Biostatistics Branch, DCEG, National Cancer Institute, 6120 Executive Blvd, 8038 Rockville, MD 20852. E-mail: joshua.sampson@nih.gov
Received 27 September 2010; Revised 24 December 2010; Accepted 12 January 2011
Published online 2 March 2011 in Wiley Online Library (wileyonlinelibrary.com/article/gepi).
DOI: 10.1002/gepi.20575

INTRODUCTION

In order to capitalize on the rapid advancement of Next Generation Sequencing (NGS) technologies [Metzker, 2010; Shendure and Ji, 2008], investigators have initiated a range of studies designed to capture the spectrum of genetic variants underlying human diseases and traits. In particular, there has been an emphasis on exploring rare variants (minor allele frequency (MAF) $<1\%$), and uncommon variants (MAF between 1 and 10%). As expected, initial reports herald the improvements in mapping classical Mendelian disorders within pedigrees and studying more complex, oligogenic diseases. So far, NGS studies have already identified DNA variants responsible for Kabuki syndrome [Ng et al., 2010b], Miller Syndrome [Ng et al., 2010a], acute myeloid leukaemia [Ley et al., 2008] and metachondromatosis [Sobreira et al., 2010]. As these findings exemplify the power of the technology, we must now consider the refinements needed to efficiently design and conduct future studies.

NGS technologies represent major advances in the scope of sequencing, but are still susceptible to platform-specific chemistry, imaging and PCR incorporation errors. As a result, the error rate per sequenced base is often as high as 1–2%. Between this high error rate and the need to identify the alleles on both chromosomes, each position must be covered by multiple reads to ensure that variation is accurately classified [Harismendy et al., 2009]. As the cost of the study will increase with the total number of reads, the minimal coverage depth necessary for identifying variant, or rare, alleles should be determined. Moreover, when the choice is between increasing coverage depth and increasing the number of individuals, it is important to determine the depth that maximizes power.

We start by examining the relationship between the average coverage depth and the proportion of rare alleles that can be detected within an individual. Since common practice is to determine average depth, as opposed to the depth at each base, the proportion detected depends heavily on the distribution of coverage across the genome. We discuss this distribution, using data from the 1000 Genomes Project [Kuehn, 2008; The 1000 Genome Project Consortium, 2010], and show that a single *shape* parameter can provide a new measure of quality for NGS technology. Our discussion is applicable to currently available calling methods [Bansal et al., 2010; Horner et al., 2010; Quinlan et al., 2008; Shen et al., 2010].

Next, we consider the goal of identifying rare variants within a sample of unrelated individuals from a given population. Using an illustrative model, in which the sequencing cost can be considered as a product of the number of subjects and the average coverage depth, we discuss the trade-off between the two. This model has simplified the cost structure by ignoring other elements such as DNA extraction, library preparation, and bioinformatic analysis, but keeps the essential point, that increasing either sample size size or depth often must be accompanied by a decrease in the other. This trade-off has also been considered by other groups, including the 1000 Genomes project [Kaiser, 2008; Kuehn, 2008; Siva, 2008], Bhangale et al. [2008] and Wendl and Wilson [2008, 2009a,b]. The

effects of changing the parameters of the study, such as read error rate, shape parameters and statistical test, are explored when trying to identify a suitable combination of sample size and coverage depth to detect variants. When designing a case/control study to identify associations between SNPs and a disease, sample size must also be balanced against coverage depth. Therefore, we offer a method for determining a combination that can maximize the power for detecting influential SNPs. This discussion extends previous work that identified the sample sizes needed to detect rare variants for association studies without considering error [Li and Leal, 2010].

METHODS

NOTATION AND ASSUMPTIONS

Let individual i have K_{ij} reads covering position j (all notation is listed in Table I). Let A_{ijk} be the true allele for read k, $k \in \{1,\ldots,K_{ij}\}$ and let $\hat{A}_{ijk}$ be the observed allele. Let $P(\hat{A}_{ij\cdot}|K_{ij}, G_{ij}=1)$ be the probability of observing a specific set of alleles given individual i is heterozygous, $G_{ij}=1$, at position j, conditioned on the number of reads. Similarly, let $P(\hat{A}_{ij\cdot}|K_{ij}, G_{ij}=0)$ be the probability given the individual has no variants, $G_{ij}=0$, at position j. Note, vectors are denoted by the '$\cdot$' in the subscript. To calculate these probabilities, we make five assumptions.

A1. The number of reads, K_{ij}, is distributed as a poisson variable with mean $\mu\lambda_j$.

A2. λ_j is distributed as a gamma variable with shape parameter ζ and scale parameter $1/\zeta$.

A3. Given an individual's genotype, the probability of observing a variant allele on a read covering j follows the bernoulli distribution with mean p_{MA} (often, $p_{\mathrm{MA}}=0.5$) when $G_{ij}=1$, r, the read error rate, when $G_{ij}=0$, and $1-r$ when $G_{ij}=2$.

$$P(\hat{A}_{ijk}=1|G_{ij}) = \begin{cases} r & \text{if } G_{ij}=0, \\ p_{MA} & \text{if } G_{ij}=1, \\ 1-r & \text{if } G_{ij}=2. \end{cases} \quad (1)$$

A4. Any two reads, even those within the same individual, are independent conditional on the genotype: $\hat{A}_{ijk_1} \perp \hat{A}_{ijk_2}|G_{ij}$.

A5. The frequencies of G_{ij} follow Hardy-Weinberg Equilibrium.

TABLE I. Notation

n	Number of subjects
	Label the subjects with $i \in \{1,\ldots,n\}$
N	Number of bases, or loci
	Label the loci with $j \in \{1,\ldots,N\}$
T	Total number of bases in all reads
Y_i	Disease status of subject i
	$Y_i \in \{0,1\}$
θ_j	Minor allele frequency (MAF) of locus j
G_{ij}	Genotype, or # of minor alleles
	$G_{ij} \in \{0,1,2\}$
K_{ij}	Number of reads containing locus j for subject I
A_{ijk}	True allele copied in read k
	$A_{ijk}=1$ denotes the minor allele
$\hat{A}_{ijk}$	Observed allele in read k
r	Probability observed and true alleles differ
	$r \equiv P(\hat{A}_{ijk} \neq A_{ijk})$
μ	Average coverage depth

DISTRIBUTION OF DEPTH OF COVERAGE

Our first goal is to find the best model fit for the distribution of read depth. Low coverage sequencing runs for Caucasian individuals from the 1000 genomes project are the first examples. BAM files from samples sequenced at the Sanger Center on the ILLUMINA platform and samples from Baylor College of Medicine on the ABI Solid platform were downloaded from ncbi.nlm.nih.gov: 1000genomes/ftp/data/ using Aspera. The pileup command in SAMTOOLS was then used to obtain the depth of coverage for the first position of all dbSNP entries on chromosome 1. Versions of the sequence and alignment indices were from 3/11/2010. For a second example, we downloaded read information for the YH genome, the first genome sequenced as part of the YanHuang project, from http://yh.genomics.org.cn.

For each sample, we fit the distribution of read depth by a negative binomial distribution. A value of ζ is selected for each platform. For samples within a platform, the MLE of the sample means, $\mu_1,\ldots,\mu_n$, are calculated based on only those positions with at least one read and the poisson assumption. We defined the shape parameter to be the value of ζ that minimized

$$\sum_i \sum_{d=1}^{100} \hat{f}_i(d)(\hat{f}_i(d) - f(d|\zeta, \hat{\mu}_i))^2, \quad (2)$$

where $f(d|\zeta,\mu_i)$ is the density of a truncated negative binomial distribution at depth d, $\hat{f}_i(d)$ was the observed density at depth d in subject i among all positions with at least one read, and 100 was chosen as an upper bound.

RARE VARIANT DETECTION WITHIN AN INDIVIDUAL

We calculate the power to detect a rare variant within an individual under scenarios where we vary r, ζ, p_{MA}, μ and α_{IND}, the allowed false-positive rate. Our method for calculating power is best described by simulation. For each scenario, we simulate λ_j for S SNPs from a gamma distribution with shape parameter ζ and scale parameter $1/\zeta$. For each SNP, we then simulate K_{ij} from a poisson distribution with mean $\mu\lambda_j$. To simplify the simulations, we could combine these steps by generating K_{ij} according to a negative binomial distribution with mean μ and size ζ. To identify the null distribution, we then simulate the number, v_{ij}, of rare variants at SNP j from a binomial distribution(K_{ij},r), where r is the known read error rate, and calculate LRS_{ij}, where LRS_{ij} is the likelihood ratio statistic comparing $G_{ij}=1$ and $G_{ij}=0$,

$$\mathrm{LRS}_{ij} = 2\log(L_{ij}) = 2(\log(0.5^{K_{ij}}) - \log(r^{v_{ij}}(1-r)^{K_{ij}-v_{ij}})), \quad (3)$$

and L_{ij} is the likelihood ratio

$$L_{ij} = \frac{P(\hat{A}_{ij\cdot}|K_{ij}, G_{ij}=1)}{P(\hat{A}_{ij\cdot}|K_{ij}, G_{ij}=0)}. \quad (4)$$

The α_{IND} threshold, t_α, is the $(1-\alpha_{\mathrm{IND}})$ quantile of the S values of LRS_{ij}. To calculate power, we simulate the number of rare variants at each SNP from a binomial (K_{ij},0.5), calculate LRS_{ij} and then define power as the proportion of values exceeding t_α. In practice, simulation could be replaced by calculating power directly (see the "Rare Variant Detection within a Population Sample" section on the next page).

RARE VARIANT DETECTION WITHIN A POPULATION SAMPLE

We calculate the power to detect a rare variant within a population under scenarios where we vary r, ζ, μ, α, θ and n. Power calculations start by simulating a set of S datasets, $S = 5{,}000{,}000$ for the null distribution and $S = 5{,}000{,}000$ for the alternative distribution. For each dataset, n values of (K_{ij}, v_{ij}) are generated. For each dataset, we calculate the likelihood ratio statistic, LRS_j^*, to test the null hypothesis $\theta = 0$, where θ is the MAF.

$$\mathrm{LRS}_j^* = 2(\ell^*(\hat{\theta}_{\mathrm{MLE}}, r|\vec{A}_{\cdot j\cdot}) - \ell^*(0, r|\vec{A}_{\cdot j\cdot})), \tag{5}$$

where

$$\ell^*(\theta|\vec{A}_{\cdot j\cdot}, r) = \sum_i \log((1-\theta)^2 r^{v_{ij}}(1-r)^{K_{ij}-v_{ij}} + \theta(1-\theta)0.5^{K_{ij}-1} + \theta^2 r^{K_{ij}-v_{ij}}(1-r)^{v_{ij}}), \tag{6}$$

and $\hat{\theta}_{\mathrm{MLE}}$ is the maximum likelihood estimator. To estimate the null distribution, each dataset is presumed to include n individuals with $G_{ij} = 0$ so v_{ij} is distributed as a binomial (K_{ij}, r) for all i. The S values of LRS_j are then calculated, with t_α defined as the $(1-\alpha)$ quantile of those S values. To calculate power, we simulate S datasets, and for each dataset, generate the genotypes of the n individuals by a multinomial distribution with probabilities $(\theta^2, 2\theta(1-\theta), (1-\theta)^2)$, and generate v_{ij} according to a binomial with parameters (K_{ij}, r) if $G_i = 0$, $(K_{ij}, 0.5)$ if $G_i = 1$, and $(K_{ij}, 1-r)$ if $G_i = 2$. The power is the proportion of the S values of LRS_j^* that exceed t_α.

DETECTING ASSOCIATIONS WITH DISEASE

We calculate the power to detect a rare variant within a population under scenarios where we vary r, ζ, μ, α, θ, n and RR, the relative risk attributable to the variant allele under the additive model. Power is calculated directly, without simulation, as follows. We assume a disease prevalence of 0.1 and calculate M_1, the 2×3 matrix containing the probabilities of each genotype in controls and cases. Then for each possible genotype, we calculate the distribution of L_{ij} as discussed in the first section of the Appendix and the resulting distribution of the called genotypes, where $\hat{G}_{ij} = 0$ if $L_{ij} < 1$, $\hat{G}_{ij} = \mathrm{NA}$ (i.e. no call) if $1 \le L_{ij} < t_1$, $\hat{G}_{ij} = 1$ if $t_1 \le L_{ij} < t_2$, and $\hat{G}_{ij} = 2$ if $\log(L_{ij}) \ge t_2$. The optimal thresholds, t_1 and t_2, require calculation and are discussed in the third section of the Appendix. Given these probabilities, we calculate the 3×4 matrix, M_2, containing the probability of $\hat{G}_{ij}$ (columns) conditioned on G_{ij} (rows). Letting the fourth column of M_2 be the conditional probabilities of a ''no-call'', the first three columns of the product $M_1 M_2$ are the expected counts for the observed genotypes in a given study. From this 2×3 table, we calculate the non-centralty parameter (ncp) for an association test for a dominant variant and define power as $P(\chi^2_{1,\mathrm{ncp}} > t_{\chi^2_1,\alpha})$ where $\chi^2_{1,\mathrm{ncp}}$ follows a χ^2 distribution and $t_{\chi^2_1,\alpha}$ is the appropriate threshold.

RESULTS

DISTRIBUTION OF COVERAGE DEPTH

Given a total of R reads, the ideal distribution spreads them uniformly over the genome so that every base is covered by the same number of reads. Both stochastic and experimental limitations prevent such uniform coverage, and specification is limited to the average, $\mu = T/N$, number of reads, where T is the total number of bases in the R reads (i.e. $R \times$ average read length) and N is the number of bases to be sequenced within an individual (all notation is listed in Table I). If the efficiency and quality of sequencing were constant across the genome, the distribution of the number of reads at any given position (i.e. K_{ij}), or depth of coverage, would likely share similar characteristics to that of a poisson variable with mean μ. Unfortunately, because of variation in local content, such as the proportion GC and extent of segment duplication, some regions in the genome can, for certain NGS technologies, be more difficult to sequence and align [Kidd et al., 2010]. Therefore, it is unlikely that the overall distribution of reads follows a poisson distribution, and evidence from our own data, full genome sequencing [Wang et al., 2008], and the 1000 genomes project suggests that the distribution appears to be most similar to a negative binomial (Fig. 1). To understand the origin of such a distribution, consider that each base, defined by its location in the genome, has its own, intrinsic, sequencing inclination λ_j and, given this value, the number of reads covering that base will follow a poisson distribution with mean $\mu\lambda_j$. If these λ_j were distributed according to a gamma variable with shape ζ and scale $1/\zeta$, then the overall read depth should be distributed as a negative binomial with mean μ and size ζ. We chose to work with the gamma distribution as it conveniently characterizes the quality of different technologies using a single parameter and seems to fit data very well.

The shape parameter (ζ), which describes how the λ_j are distributed, is another important measure of sequencing performance. With the goal being consistent performance across the genome, all values of λ_j would ideally be identical and equal to 1. As larger ζ imply more tightly distributed λ_j, better technologies will, in general, have larger ζ, a quantity that can be estimated by the data. Specifically, for each individual, we fit the observed distribution of coverage depth to a truncated negative binomial distribution that considers only positions with at least one read. Therefore, interpretation of ζ as a measure of quality is only meaningful when also considering the total number of reads and the total number of loci covered.

We estimated ζ values from four sets of samples. Two sets of samples were from the 1000 genomes project. For the distributions of reads corresponding to variants in dbSNP along chromosome 1 in 116 samples analyzed with the ILLUMINA platform at the Sanger Center, we estimated $\zeta = 4$ (Fig. 1, see supplementary material for all 116 fits). For the distributions from 20 samples analyzed by the ABI SOLID platform at the Baylor College of Medicine, we estimated $\zeta = 5.5$. In addition to platform, study center also affected ζ as the shape parameter appears slightly higher from data collected on the ILLUMINA platform at ILLUMINA. Furthermore, we examined the distribution from a study with higher coverage. For the distribution of reads along chromosome 1 in the YH individual [Wang et al., 2008] sequenced in 2008 by Illumina Genome Analysers, we estimate $\zeta = 7$. Finally, we estimated ζ in an exome sequencing project (unpublished data), where base performance now depends on the ease of sequencing and ease of capture. Because the capture step is far more heterogenous, it was not surprising to find $\zeta = 2.2$. As

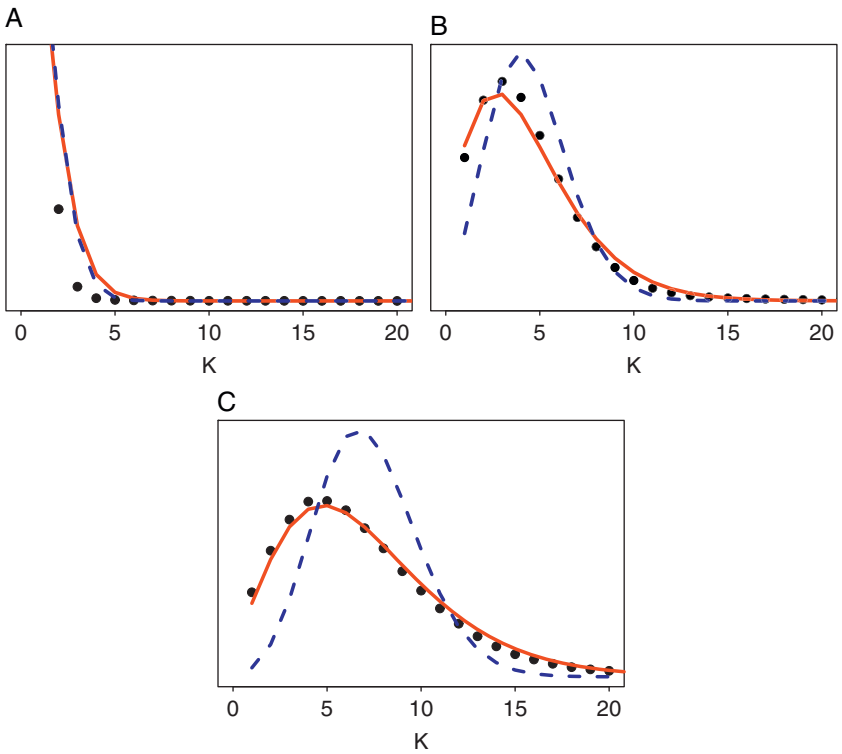

Fig. 1. Comparing the observed distribution of coverage depth to the distribution estimated by the model. Graphs are for the three subjects, with the lowest, median, and highest, coverage depth, among the 116 individuals genotyped at the Sanger Center. The black dots show the true proportion of SNPs with the specified read depth (*x*-axis), the red/unbroken line shows the distribution of a negative binomial with that subject's mean depth and $\zeta = 4$, and the blue/dashed line shows the poisson distribution.

Figure 2 shows that such a low ζ would noticeably reduce power to detect rare variants. As most upcoming studies focus on targeted regions, these studies need to expect small values of ζ and should choose their depth accordingly.

RARE VARIANT DETECTION WITHIN AN INDIVIDUAL

Consider testing whether an individual has a variant allele at a given location. In hypothesis testing language, our null hypothesis will be that the individual does not have a variant allele. Power (pow_{IND}) is then defined to be the probability of correctly rejecting H_0 when the subject is truly heterozygous ($G_{ij} = 1$). Power at a single location and percentage of variant alleles that are discovered are identical quantities. The false-positive rate, α_{IND}, is the probability of incorrectly rejecting H_0 when $G_{ij} = 0$.

Figure 2A shows the power to detect a heterozygote as a function of μ. If the desired false-positive rate is $\alpha_{\text{IND}} = 10^{-5}$ (i.e. 1/10,000 homozygous positions is misclassified as heterozygous), then the proportion of variant alleles expected to be discovered across the genome are 80, 90, 95 and 99% when the average depth is 12, 17, 23 and 42, respectively. If the acceptable false-positive rate is lowered to $\alpha_{\text{IND}} = 10^{-6}$, then the required depths would be 14, 20, 27 and 49. If the acceptable false-positive rate is raised to $\alpha_{\text{IND}} = 10^{-4}$, the required depths would be 10, 14, 19 and 25. The cost, in depth, of raising power by 1% increases dramatically with power.

Figure 2B shows the dependence of power on the shape parameter, ζ. As the shape parameter shrinks, and the distribution of K_{ij} diverges farther from poisson, the mean depth needed to ensure an acceptable level of power will increase. When λ_j follows a poisson distribution, a depth of 15 detects 95% of the variant alleles ($\alpha_{\text{IND}} = 10^{-5}$), whereas a depth of 23 is needed when $\zeta = 4$. A depth of $\mu = 15$, with $\zeta = 4$, only detects 85% of the variant alleles. For exome-sequencing, with $\zeta = 2.2$, a depth of $\mu = 15$ only detects 79% of the variant alleles.

Furthermore, there is concern that the variant allele may be more difficult to copy and therefore occur at a lower frequency (i.e. p_{MA}, defined in assumption 3 in the earlier section, may be below 0.5). If only 40% of the reads at a heterozygous position are expected to be the variant allele, the proportion detected shrinks by ~6% ($\alpha_{\text{IND}} = 10^{-5}$, $\zeta = 4$, $\mu = 23$: 89 vs. 95%). Figure 2C shows that, if true, read bias can result in a large loss in power.

RARE VARIANT DETECTION WITHIN A POPULATION SAMPLE

Consider testing the null hypothesis that all individuals are homozygous at locus j, $G_{\cdot j} = \vec{0}$, in a random sample of unrelated individuals from a given population. We consider the likelihood ratio test, defined in the earlier section, as it performs better than tests based on individuals' assigned genotypes.

The power to detect a rare allele increases with n as the likelihood of collecting an individual with a rare-variant

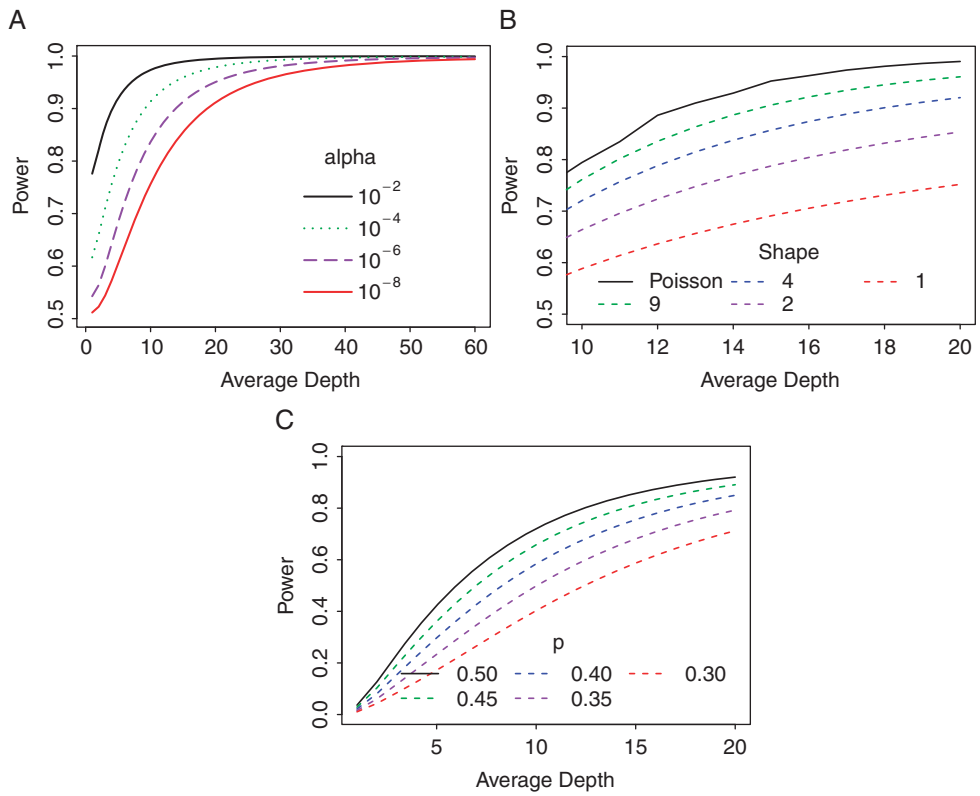

Fig. 2. (A) The power to detect a heterozygote individual as a function of average depth and α-level when $r = 0.01$, $p_{MA} = 0.5$, and assuming K_{ij} follows a negative binomial distribution. (B) The power to detect a heterozygote individual for different values of ζ ($\alpha = 10^{-5}$). Note the change in *x*-axis. (C) The power to detect a heterozygote individual for different values of p_{MA}, or different read biases ($\alpha = 10^{-5}$, $\zeta = 4$).

increases with n. For any given MAF, however, there is an n at which the power plateaus as the chance that none of the subjects have the rare variant is negligible. Power also increases with coverage depth and for any given MAF, there will be a μ for which the likelihood of incorrectly classifying a true heterozygote is negligible. Figure 3A illustrates power as a function of μ and n when MAF = 0.005, $r = 0.001$, $\alpha = 0.0001$ and λ follows a poisson distribution. Darker colors indicate higher probabilities of detection.

If we fix the total number of reads, letting $\mu_{big} = \mu \times n$, our proxy for cost, there will be a specific combination of n and μ that maximize the number of rare variants detected. In the above example, setting $\mu_{big} = 500$, we could choose any combination along the 500/μ contour (Fig. 3A). To better illustrate the options, we plot power as a function of μ with the total cost fixed (Fig. 3B). Note, similar to previous findings [Wendl and Wilson, 2009a,b], choosing a λ less than the optimal value reduces power far more than choosing a λ greater than the optimal value.

As illustrated in Figure 3C, the optimal λ depends most strongly on r, and greater depth is required as the read error rate increases. When we reduce our tolerance for false positives (i.e. decrease α), we similarly benefit from increasing depth. Changing other parameters has less effect on the optimal λ, as previously suggested in Wendl and Wilson [2009a,b]. Reducing the shape parameter and/or increasing μ_{big} may slightly reduce the optimal μ. Overall, depending on the parameters of the experiment, we find the optimal μ to be between 2 and 8.

DETECTING ASSOCIATIONS WITH DISEASE

Consider testing for the association between a rare or uncommon variant and a disease in a case/control study with individuals split equally between the two groups. Here, we present results from a genotype-based test, where we first genotype each individual and then perform a χ^2 test of association, a standard test for the case/control analysis. In contrast to a test for rare variants, the association test will have maximum power if n is as large as possible when the total cost, as defined by μ_{big}, is fixed (Fig. 4). We review the intuition for this observation in the discussion section. In this section, we examine the rate at which increasing coverage decreases power. The decrease in power from raising $\mu = 1$ to $\mu = 2$, with the corresponding decrease in sample size, primarily depends on the overall power of the association test when $\mu = 1$. Power is extremely sensitive to changes in sample size. As a specific example, consider the scenario where the total number of reads is 20,000, $r = 0.01$, MAF = 0.05, and $\alpha = 0.0001$. Note, MAF is set at a higher frequency than previous examples to achieve detectable power. By varying the relative risk from 1.2 to 2.0, we vary the power at $\mu = 1$ from 0.03 to 0.999. For relative risks of 1.4, 1.6 and 1.8, increasing μ from 1 to 2 decreases the power from 0.42, 0.88 and 0.99 to

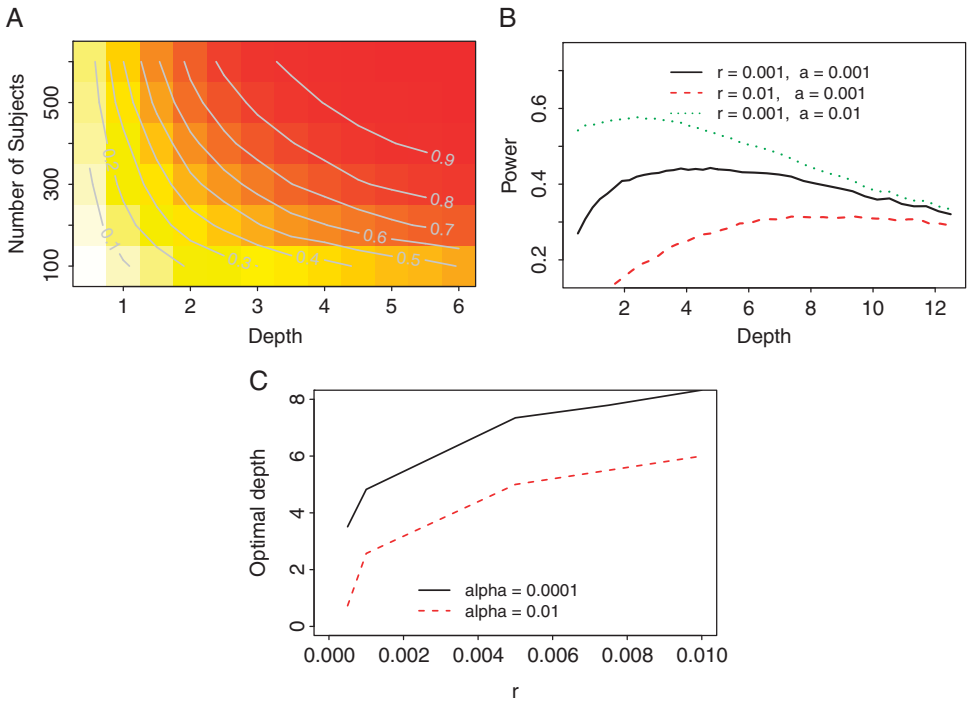

Fig. 3. *Main Figure*: **(A) The power to detect a rare variant for all possible combinations of n (number of subjects) and μ (depth of sequencing) using a likelihood ratio test, with $\lambda_j \sim$Poisson, $r = 0.001$, $a \equiv \alpha = 0.0001$, and MAF = 0.005. (B) The black/unbroken line is power as a function of μ when $n \times \mu = 500$ with the above parameters. The red/dashed and green/dotted lines show the μ/power relationships when $r = 0.01$ and $a \equiv \alpha = 0.01$, respectively. (C) The black/unbroken line shows the optimal λ as a function of read error rate, with $\alpha = 0.0001$, MAF = 0.005, $n \times \mu = 500$, and λ∼Poisson. The red/dashed line shows the optimal λ when α is raised to 0.01.**

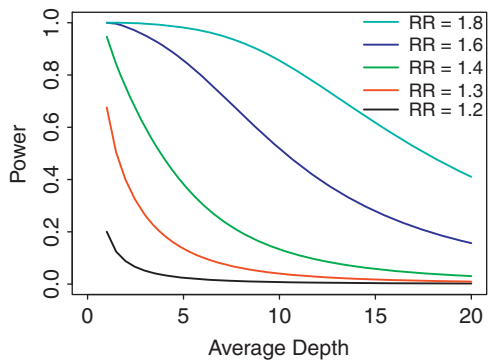

Fig. 4. For fixed cost, the power to detect an association decreases with μ. The sharpness of the decline depends on the relative risk attributable to the SNP. The relationship between power and μ is illustrated for four different values of the relative risk when MAF = 0.05, $r = 0.01$, $\alpha = 0.0001$, and $n \times \mu = 50{,}000$.

0.20, 0.64 and 0.90. Note that the loss in power can be relatively large for even small increases in μ when the power is low. Holding power constant, changing other variables such as the total number of reads, MAF or r, has little, if any, effect on the rate of decrease. Changing α only has a slight effect, in that allowing for a higher false-positive rate dampens the rate of decrease.

Again, the genotyping test is not the most powerful test and the preferable test would be the likelihood ratio test. Here, the general performance of the tests are similar. In contrast to rare variant detection, the advantage of the likelihood ratio test appears truly minimal. If we consider the scenario where each base is covered by exactly one read, the likelihood ratio test and genotype-based are essentially equivalent, with both comparing the proportion of individuals with a variant allele in the two groups. For details of the likelihood ratio statistic, see the methods section. Note that this discussion assumed sequencing was performed solely for the purposes of an association test with a specified outcome and no covariates, so there was no interest in identifying any individual's specific genotype.

DISCUSSION

Next Generation Sequencing has emerged as a powerful tool for detecting rare variants and their associations with human diseases and traits. In this manuscript, we discussed how to choose the coverage depth for a study using NGS technology. Many of the statistical techniques used here were originally developed to examine the extent of overlap and coverage needed when genomes were being mapped by "fingerprinting," where overlapping clones from recombinant libraries were needed to piece together the genome [Lander and Waterman, 1988; Siegel et al., 2000; Wendl and Barbazuk, 2005; Wendl and Waterston, 2002]. We started by showing that the depth of coverage ranged greatly across the genome, especially

when performing targeted sequencing. Therefore, even when the average depth is high, a large number of positions can still have relatively low coverage. To avoid sparse coverage in difficult-to-sequence regions, we need to raise the average above that suggested by the simpler models. Statistically, we suggested that the depth of coverage followed a negative binomial distribution, as opposed to the simpler poisson distribution. A measure of the extent of deviation from the poisson ideal provides a measure of the quality of the sequencing. In this regard, the shape parameter, which describes that deviation, is an important characteristic to consider when deciding between NGS technologies.

The coverage depth needed to identify a variant allele with high specificity was surprisingly large. The exact choice of depth depends on multiple considerations, such as desired α-level, quality of technology and read error rate. If we are only looking for variants at a pre-specified set of loci, perhaps those positions already known to be polymorphic, we might allow a relatively large α-level, 1/1000 or even 1/100. At such rates, a coverage depth around 10 might be sufficient when using the best technologies. Another consideration is the quality of the technology. When the shape parameter is small and depth is highly variable, we would need to increase coverage depth. Our analyses were for stand-alone calling algorithms. Other algorithms, specifically those that consider linkage disequilibrium, will likely require lower depth. Results from the 1000 genome data should help show how such advanced methods can augment power.

When trying to detect a new rare variant within a population, the desirable coverage depth, with realistic parameters, ranges between 2 and 8 reads, with the exact depth depending heavily on the acceptable false-positive rate. With too few reads, it could be difficult to determine whether a few variant alleles, scattered across all subjects, occurred by error. With too few individuals, there would be a non-negligible chance that none of the individuals carried the rare variant. Therefore, the depths that perform well balance between these extremes. Our suggested coverage depth is similar to the depth, $4-6\times$, chosen for the 1000 Genomes project and Wendl's approximation of 3.6.

In contrast with the test for rare variants, the association test is maximized for power by including as many cases as possible. In studies with secondary analyses, adjustment for confounders, or a planned follow-up, increasing depth to identify heterozygous individuals may be necessary. However, the consequences of such increases can be severe, if they require a reduction in sample size. Li and Leal also discuss the need for large sample sizes when detecting rare haplotypes with frequency $<1\%$ [Li and Leal, 2008].

An association test requires genotype calling optimized for that purpose. Although it is an extreme example, consider using a calling algorithm that requires 50 reads before making a call. Obviously, with such a rule, using a low average depth will perform poorly. Therefore, for a fair comparison, we should use the optimal calling rule, or alternatively a rule which maximizes the power of the following association test. As no such rule had been discussed previously in the literature, we derived it in a general, and broadly applicable, form in the appendix.

To understand why maximizing the number of subjects is optimal for association studies, consider the simple case where there are no read errors and all SNPs with at least one read of the minor allele are called as $\hat{G}_{ij}=1$. As a standard rule, power is determined by the number of events, which, in this case, is the number of called heterozygotes. Note that $\sum \hat{G}_i$ is higher for $2n$ subjects with 1 read per base compared to n subjects with 2 reads per base. Let p be the true proportion of heterozygotes. When the reads for only 50% of the $G_i=1$ individuals contain a minor allele, we would expect $2n\times p\times 0.5$ individuals to have $\hat{G}_{ij}=1$. In the alternative, where 75% of those individuals will have at least one read with a minor allele, we would expect $n\times p\times 0.75$ to have $\hat{G}_{ij}=1$. Note that the gain accuracy does not offset the loss in the number of subjects, $2n\times p\times 0.5>2n\times p\times 0.75$. See the supplementary material for addition discussion.

There are important limitations for our conclusions. First, we assume a constant error rate, regardless of MAF or the number of minor alleles detected. Because calling algorithms often use linkage disequilibrium to aid calling, per base error rates can decrease as sample increases. One of the consequences is that the depth needed to detect a heterozygous allele within an individual actually depends on the total sample size. Second, we have focused only on SNPs. Because of an increased difficulty in detecting structural variants, such as copy number variation, insertions, inversions and translocations [Feuk et al., 2006], the optimal depth for detecting this type of variation may require a depth greater than 50 when using the discordant read pairs method [Wendl and Wilson, 2009a,b]. However, as technology improves, reads become longer, and read error rates decrease, the necessary depth should decrease. Third, when considering the error rate at a single location, our model assumes that all errors produce the same allele. If errors were random, then the potential for observing enough of any single allele to call a specific variant allele would decrease. Fourth, we only take into consideration sequencing costs. Other costs, such as collecting data and sample preparation, have been ignored from our cost structure.

As the cost of NGS falls, its use will continue to increase, and thus it will be necessary to optimize designs to efficiently discover and validate variants that map to human diseases and traits. Sequencing with too low a depth can negatively impact discovery and with too few individuals can negatively impact detecting associations. Overall, it is important to balance sample size and coverage depth in the context of available resources for NGS studies.

ACKNOWLEDGMENTS

We greatly appreciate the Genetic and Epidemiology Branch, DCEG, NCI for allowing us access to the data from their exome studies. K.J., M.Y. and S.C. identified the problem. J.S. and N.C. developed the statistical framework. J.S. ran simulations and drafted the paper. All authors revised the paper.

REFERENCES

Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA. 2010. Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res 20:537–545.

Bhangale TR, Rieder MJ, Nickerson DA. 2008. Estimating coverage and power for genetic association studies using near-complete variation data. Nat Genet 40:841–843.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7:85–97.
Harismendy O, Ng P, Strausberg R, Wang X, Stockwell T, Beeson K, Schork N, Murray S, Topol E, Levy S, Frazer K. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10:R32.
Horner DS, Pavesi G, Castrignano T, De Meo PD, Liuni S, Sammeth M, Picardi E, Pesole G. 2010. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform 11:181–197.
Kaiser J. 2008. A plan to capture human diversity in 1000 genomes. Science 319:395.
Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, Alkan C, Malig M, Ventura M, Giannuzzi G, Kallicki J, Anderson P, Tsalenko A, Yamada NA, Tsang P, Kaul R, Wilson RK, Bruhn L, Eichler EE. 2010. Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Meth 7:365–371.
Kuehn B. 2008. 1000 genomes project promises closer look at variation in human genome. JAMA 300:2715.
Lander E, Waterman M. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–239.
Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling DD, Dunford-Shore BH, McGrath SM, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott SA, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla AC, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich JI, Heath SH, Shannon WDS, Nagarajan RN, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK. 2008. DNA sequencing of a cytogenetically normal acute myseloid leukaemia genome. Nature 456:66–69.
Li B, Leal S. 2008. Discovery of rare variants via sequencing: implications for association studies [abstract]. Genet Epidemiol 32:702.
Li B, Leal SM. 2010. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet 5:e1000481.
Metzker M. 2010. Sequencing technologies—the next generation. Nat Rev Genet 1:31–46.
Ng SB, Buckingham KJ, Lee C, Bigham A, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad M. 2010a. Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 42:30–35.
Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K-i, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. 2010b. Exome sequencing identifies mll2 mutations as a cause of kabuki syndrome. Nat Genet, advance online publication.
Quinlan AR, Stewart DA, Stromberg MP, Marth GT. 2008. Pyrobayes: an improved base caller for snp discovery in pyrosequences. Nat Meth 5:179–181.
Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, Yu F. 2010. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 20:273–280.
Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat Biotech 26:1135–1145.
Siegel AF, van den Engh G, Hood L, Trask B, Roach JC. 2000. Modeling the feasibility of whole genome shotgun sequencing using a pairwise end strategy. Genomics 68:237–246.
Siva N. 2008. 1000 genomes project. Nat Biotechnol 26:256.
Sobreira NLM, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, Gumbs CE, Pevsner J, Thomas G, Valle D, Hoover-Fong JE, Goldstein DB. 2010. Whole-genome sequencing of a single proband together with linkage analysis identifies a mendelian disease gene. PLoS Genet 6:e1000991.
The 1000 Genome Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.
Wang J, Wang J, Li W, Li R, Tian Y, Goodman G, Fan L, Zhang W, Li J, Zhang J, Guo J, Feng Y, Li B, Lu H, Fang Y, Liang X, Du H, Li Z, Zhao D, Hu Y, Yang Y, Zheng Z, Hellmann H, Inouye I, Pool M, Yi J, Zhao X, Duan J, Zhou J, Qin Y, Ma J, Li L, Yang G, Zhang Z, Yang G, Yu B, Liang C, Li F, Li W, Li S, Ni D, Ruan P, Li J, Zhu Q, Liu H, Lu D, Li Z, Guo N, Zhang G, Ye J, Fang J, Hao L, Chen Q, Liang Q, Su Y, San Y, Ping A, Yang C, Chen S, Li F, Zhou L, Zheng K, Ren H, Yang Y, Gao L, Yang Y, Li G, Feng Z, Kristiansen X, Wong K, Nielsen GKS, Durbin R, Bolund R, Zhang L, Li X, Yang S, Wang H, Jian. 2008. The diploid genome sequence of an Asian individual. Nature 456:60–65.
Wendl M, Barbazuk W. 2005. Extension of Lander-Waterman theory for sequencing filtered DNA libraries. BMC Bioinform 6:245.
Wendl M, Waterston R. 2002. Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. Genome Res 12:1943–1949.
Wendl M, Wilson R. 2008. Aspects of coverage in medical DNA sequencing. BMC Bioinform 9:239.
Wendl M, Wilson R. 2009a. Statistical aspects of discerning indel-type structural variation via DNA sequence alignment. BMC Genomics 10:359.
Wendl M, Wilson R. 2009b. The theory of discovering rare variants via DNA sequencing. BMC Genomics 10:485.

APPENDIX

RARE VARIANT DETECTION WITHIN AN INDIVIDUAL

We calculate power, β, using a numerical approximation of

$$\beta = \int_{\lambda_j} \left[\sum_{K_{ij}=1}^{\infty} \left(\sum_{v_{ij}=0}^{K_{ij}} P(K_{ij}|\mu\lambda_j) P\left(\sum_k \hat{A}_{ijk} = v_{ij} | G_{ij} = 1 \right) 1(L_{ij} \geq t_\alpha) \right) \right] \times dF(\lambda_j), \quad \text{(A1)}$$

where we choose t_α so

$$P(\ell_{ij} > t_\alpha | \lambda_j) = \sum_{K_{ij}=1}^{\infty} \left(\sum_{v_{ij}=0}^{K_{ij}} P(K_{ij}|\mu\lambda_j) P\left(\sum_k \hat{A}_{ijk} = v_{ij} | G_{ij} = 0 \right) 1(L_{ij} \geq t_\alpha) \right) = \alpha_{\text{IND}}. \quad \text{(A2)}$$

TESTING FOR ASSOCIATION: LIKELIHOOD RATIO TEST

We need only find $\hat{\theta}_U$, $\hat{\theta}_A$ and $\hat{\theta}_C$, those estimates that maximize the log-likelihood for controls alone, cases alone and cases/controls combined. The log likelihood for each group can be defined by summing over their respective members

$$\ell_X(\theta | r, \vec{A}_{\cdot j}) = \sum_{i \in X} \log((1-\theta)^2 r^{v_{ij}} (1-r)^{K_{ij}-v_{ij}} + \theta(1-\theta) 0.5^{K_{ij}-1} + \theta^2 r^{K_{ij}-v_{ij}} (1-r)^{v_{ij}}) \quad \text{(A3)}$$

and the likelihood ratio statistic

$$\text{LRS}_j^* = 2(\ell_U(\hat{\theta}_U | r, \vec{A}_{\cdot j}) + \ell_A(\hat{\theta}_A | r, \vec{A}_{\cdot j}) - \ell_C(\hat{\theta}_C, r | \vec{A}_{\cdot j})). \quad \text{(A4)}$$

TESTING FOR ASSOCIATION: OPTIMAL THRESHOLD

When testing for association, the true difficulty is determining the optimal calling algorithm when genotyping individuals. Here, optimal implies maximizing the power of the upcoming association test. Although the threshold for calling $\hat{G}_{ij} = 0$ can be set to 1 without worry, t_1 requires more consideration. We ignore t_2 because we test association under the dominant model. If t_1 is too large, then we can discard a substantial proportion of true heterozygotes. If t_1 is too small, then we can misclassify too many homozygotes. Either error reduces power. By the calculation below, the optimal threshold, t_1 is the smallest value of t satisfying

$$\text{MAF} \geq 0.5\frac{P(L_{ij} = t|G_{ij} = 0)}{P(L_{ij} = t|G_{ij} = 1)} - \frac{P(L_{ij} \geq t|G_{ij} = 1)}{P(L_{ij} \geq t|G_{ij} = 0)}. \quad \text{(A5)}$$

Equation (A5) shows two key properties of the optimal threshold

1. The optimal threshold will be higher for alleles with low MAF, as the right-hand side of Equation (A5) is decreasing with t. When the MAF is low, we need to keep t_1 high to prevent false positives from overwhelming those few true heterozygotes.
2. The optimal threshold will be higher when the error rate is high. When the read error is high or μ is low, we need to be extra vigilant about preventing false positives.

Equation (A5) is based on performing a χ^2 test on a 2×2 contingency table. This discussion is completely general, and this threshold will be optimal whenever a continuous variable is being converted to a 0/1 variable. In the following 2×2 contingency table, let a, b, c and d be the expected cell counts.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Then, the ncp from the χ^2 test for equal proportions can be written as

$$\eta \equiv \frac{(ac - bd)^2}{(a+b)(c+d)(a+c)(b+d)}. \quad \text{(A6)}$$

When we allow for errors and no-calls, the expected cell counts, and consequently the ncp, change. Assume, we have adopted the policy to call any locus with a $L \leq 1$ to be "AA" (i.e. $\hat{G}_{ij} = 0$), any locus with $t \leq L$ to be "AB" (i.e. $\hat{G}_{ij} = 1$) and not to call a locus with $1 < L < t$. To simplify notation we drop the ij subscript. Then, the expected cell counts are

$$\begin{bmatrix} a - aP(L>1|AA) & bP(L \geq t|AB)+aP(L \geq t|AA) \\ c - cP(L>1|AA) & dP(L \geq t|AB)+cP(L \geq t|AA) \end{bmatrix}.$$

For our calculations, we use the following notation

$$k \equiv P(L \geq t|AB), \quad \text{(A7)}$$

$$u \equiv P(L \geq t|AA). \quad \text{(A8)}$$

Therefore, the expected cell counts can be abbreviated as

$$\begin{bmatrix} a - aP(L>1|AA) & ak+bu \\ c - cP(L>1|AA) & ck+du \end{bmatrix}.$$

We can greatly simplify calculating the ncp if we make the following approximations which we believe to be reasonable

1. $a - aP(L>1|AA) \approx a$,
2. $c - cP(L>1|AA) \approx c$,
3. $a - aP(L>1|AA)+ak+bu \approx a+b$,
4. $c - cP(L>1|AA)+ck+du \approx c+d$.

Then, the ncp can be approximated by

$$\eta \approx \frac{(ad - bc)^2u^2}{(a+c)^2(a+b)(c+d)k+(a+c)(a+b)(c+d)(b+d)u}. \quad \text{(A9)}$$

By the chain rule, we know

$$\frac{d\eta}{dt} = \frac{d\eta}{dk}\frac{dk}{dt} + \frac{d\eta}{du}\frac{du}{dt}. \quad \text{(A10)}$$

Therefore, for the derivative to be less than 0, we need

$$\frac{\frac{d\eta}{du}}{-\frac{d\eta}{dk}} < \frac{\frac{dk}{dt}}{\frac{du}{dt}}. \quad \text{(A11)}$$

Straightforward calculations yield

$$\frac{du}{dt} = P(L = t|AB), \quad \text{(A12)}$$

$$\frac{dk}{dt} = P(L = t|AA), \quad \text{(A13)}$$

$$\frac{\frac{d\eta}{du}}{-\frac{d\eta}{dk}} = 2\frac{P(L \geq t|AB)}{P(L \geq t|AA)} + \frac{b+d}{a+c} \approx 2\left(\frac{P(L \geq t|AB)}{P(L \geq t|AA)} + \text{MAF}\right). \quad \text{(A14)}$$

Then we should continue to increase the threshold so long as

$$\text{MAF} < 0.5\frac{P(L = t|AA)}{P(L = t|AB)} - \frac{P(L \geq t|AB)}{P(L \geq t|AA)}. \quad \text{(A15)}$$