Briefings in Bioinformatics, 19(6), 2018, 1247–1255

doi: 10.1093/bib/bbx061
Advance Access Publication Date: 9 June 2017
Paper

Power and sample size calculations for high-throughput sequencing-based experiments

Chung-I Li, David C. Samuels, Ying-Yong Zhao, Yu Shyr and Yan Guo

Corresponding author. Yan Guo, 2220 Pierce Ave, 571 Preston Research Building, Nashville TN, 37212. Tel.: 615-445-0018; Fax: 615-936-2602; E-mail: yanguo1978@gmail.com

Abstract

Power/sample size (power) analysis estimates the likelihood of successfully finding the statistical significance in a data set. There has been a growing recognition of the importance of power analysis in the proper design of experiments. Power analysis is complex, yet necessary for the success of large studies. It is important to design a study that produces statistically accurate and reliable results. Power computation methods have been well established for both microarray-based gene expression studies and genotyping microarray-based genome-wide association studies. High-throughput sequencing (HTS) has greatly enhanced our ability to conduct biomedical studies at the highest possible resolution (per nucleotide). However, the complexity of power computations is much greater for sequencing data than for the simpler genotyping array data. Research on methods of power computations for HTS-based studies has been recently conducted but is not yet well known or widely used. In this article, we describe the power computation methods that are currently available for a range of HTS-based studies, including DNA sequencing, RNA-sequencing, microbiome sequencing and chromatin immunoprecipitation sequencing. Most importantly, we review the methods of power analysis for several types of sequencing data and guide the reader to the relevant methods for each data type.

Key words: high-throughput sequencing; power; sample size

Introduction

Recent advancement in high-throughput sequencing (HTS) technology has stimulated a range of new possibilities for biomedical research. At the same time, these advances have introduced a series of bioinformatics challenges including quality control, data storage and complexity in data analyses. Power analysis is often one of the overlooked aspects of HTS data analysis.

Power calculation is the first step in designing a successful study. Its importance is reflected by its role as the non-optional component in National Institute of Health funding applications. Sample size and power analysis have been well established for traditional biological studies, such as genome-wide association study (GWAS) and microarray gene expression studies. Compared with power analysis in GWAS and microarray gene expression studies, power analysis for HTS data-based experiments is more complicated for two major reasons. The first reason regards the unique parameters for HTS read depth and read dispersion that directly affect the ability to detect variants or gene expression, and thus need to be considered in the power analysis. Second, the number of possible applications for HTS greatly exceeds the number for microarray, introducing a variation of unique statistical scenarios for power analysis.

The most common method for categorizing HTS is by the target sequencing source (DNA versus RNA) and the analysis goal such as DNA-seq [1] (exome whole genome), RNA sequencing (RNA-seq) [2]

Chung-I Li is an assistant professor at Department of Statistics, National Cheng Kung University in Taiwan. His research focus is quality control and power analysis.
David C. Samuels is an associate professor at Department of Molecular Physiology and Biophysics, Vanderbilt University, USA. His research focus is genetics and mitochondria.
Ying-Yong Zhao is a professor at School of Life Sciences, Northwest University, China. His research is focused on genomic and epigenomic data analysis.
Yu Shyr is a professor at Department of Biostatistics, Vanderbilt University, USA. His research is focused on biostatistics methodology development including sample size and power estimation.
Yan Guo is an assistant professor in the Department of Cancer Biology, Vanderbilt University. He is also the technical director of Bioinformatics for Vanderbilt Technologies for Advanced Genomics Analysis and Research Design.
Submitted: 17 January 2017; **Received (in revised form):** 5 May 2017

(messenger RNA and total RNA), chromatin immunoprecipitation sequencing (ChIP-seq), methylation sequencing (bisulfite sequencing) [3], microbiome sequencing (16S ribosomal RNA sequencing) [4], GRO (Genomic Run-on or nuclear run on)-seq [5], cross-linking immunoprecipitation sequencing (CLIP-seq) [6], photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation sequencing [7], etc. Each type of sequencing is designed to examine a completely different problem, and often the data follow distinctly different distributions, thus requiring a specific strategy for computing power. Simply put, the power of a study is the probability of successfully detecting a given effect size. Traditional power analysis estimates the power from a given sample size, effect size and required level of statistical significance. In HTS, other factors unique to HTS data such as sequencing depth and dispersion play significant roles in determining the true power, adding an unfamiliar layer of complexity to the analysis.

It is vital to not let the excitement and challenges of HTS data overshadow the importance of power analysis. Given the high diversity of HTS applications and the complexity of power analyses for these applications, we provide a detailed review of the current status of power analysis for all major types of HTS applications, as well as recommendations for the appropriate approach to deal with power analysis in different types of study design scenarios.

DNA

Exome sequencing examines the exonic regions of the genome. Other types of DNA sequencing commonly used include whole-genome sequencing, mitochondrial DNA sequencing and other types of targeted region sequencing. The immediate goal of DNA sequencing is to identify variants such as single-nucleotide polymorphisms (SNPs), somatic mutations, insertion/deletions (indels) and structural variants (translocations, inversions, etc.) The end goal of DNA sequencing is usually to carry out a variant–phenotype association or to estimate variant frequencies in a given population. Sometimes, DNA sequencing is performed just to confirm the existence of certain variants in a few special samples or estimate the population variant frequency.

For variant–phenotype association studies, the goal is the same as in GWAS, and traditional power analysis for GWAS will apply the assumption that all variants have been inferred correctly. Power analysis for GWAS is a well-established field [8–10]. The goal for GWAS variant–phenotype association studies is to determine whether there is a statistically significant difference for the frequency of an allele between a case and a control population. The common parameters required to compute power in this situation are sample size N, effect size ρ (often stated as an odds ratio), disease prevalence and allele frequency. These types of power analysis are still relevant for association studies of common variants derived from HTS methods. However, for HTS, extra complexity is introduced with sequencing depth and read dispersion, which directly affect the probability of correctly identifying a variant, introducing a series of additional power analysis methods that we review below.

Power to detect a heterozygous variant

In traditional GWAS, SNPs are detected using genotyping arrays by clustering algorithms based on fluorescent intensity data. Traditional GWAS power analysis has been well established [11–14]. These power analyses are based on collected SNP data, and do not model the process of detecting SNPs. The goals for the HTS experiment are not limited to GWAS; for example, detecting a heterozygous variant can be the intent of the study. Currently, there is no dedicated power analysis tool for the detection of heterozygous variant using HTS data. However, for HTS, the probability of detecting an allele A (or allele B) at a given diploid genomic position follows a binomial distribution: $Binomial(D,\ p)$, where D is the depth, and P is the probability of allele A after sequencing one read that is 0.5 for all diploid genomic regions. For all heterozygous germline variants, a read has a 50% chance to represent one of the two alleles (Figure 1). In a simplified scenario, the power of detecting an alternative allele can be modeled using the binomial distribution: $Binomial(D,\ p)$.

$$\Phi\left(\frac{p-p_0}{\sqrt{p_0(1-p_0)/n}}-z_{1-\alpha/2}\right)+\Phi\left(-\frac{p-p_0}{\sqrt{p_0(1-p_0)/n}}-z_{1-\alpha/2}\right),$$

where n is the sample size, p_0 is the true probability of the alternative allele, p is the observed alternative allele frequency, $\Phi(.)$ is the standard normal distribution function, $z_{1-\alpha/2}$ is the 100(1-α/2) percentile of the standard normal distribution and α is the type I error rate [22].

However, in practice, other minor details, such as reference preferential bias, need to be considered carefully. Reference preferential bias is a form of bias that occurs when the aligner penalizes a read's alignment score when that read has a mismatch compared with the reference; this, in turn, causes the alignment score to slightly favor toward the reference allele. In HTS data, reference preferential bias can lower the observed alternative allele frequency to a range of 0.48–0.5 [23]. This reference preferential bias is usually not considered when modeling the power of SNP detection. Furthermore, the binomial distribution $Binomial(D,\ p)$ merely demonstrates the probability of detecting the alternative allele. By detecting, we mean finding a single read that supports the alternative allele, which in many cases could be the result of noise or error from the library preparation, sequencing or alignment process. Many variant callers will only call heterozygous variants after certain fixed number or a fraction of reads support the alternative allele. Some callers such as the Genome Analysis Tool Kit's [24] variant caller used a Bayesian approach by considering prior information in Single Nucleotide Polymorphism database (dbSNP) to adjust their variant calling. The power associated with different variant callers may vary depending on the exact methodology applied.

Power to detect somatic mutations and mutation frequency

Cancer treatment often benefits from knowing the expected mutation frequency in a certain gene in the patient population. Mutation frequencies have been used to guide targeted therapy in cancer treatment [25, 26]. Unlike SNPs, which are germline mutations, somatic mutations may be acquired at any time. To truly identify a somatic mutation, tumor samples need to be compared with a reference sample. Blood is usually considered the best reference [27], with the obvious exception of blood cancers. The power to identify a somatic mutation involves considering both the reference and tumor samples. The expected allele frequency for a somatically mutated allele still follows a binomial distribution: $Binomial(D,\ p)$, where D is the depth; however, the expected mutation percentage is no longer 0.5, as the tumor purity varies by sample. Studies [28, 29] have shown that the read depth ratio (reference allele versus mutated allele) can be used as an estimation of the tumor purity, which can be obtained by conducting a simulation study or using existing, similar public data. However, the mutation percentage at each genomic position might also vary, which makes modeling the power for somatic mutation detection difficult. The power to detect somatic mutations is

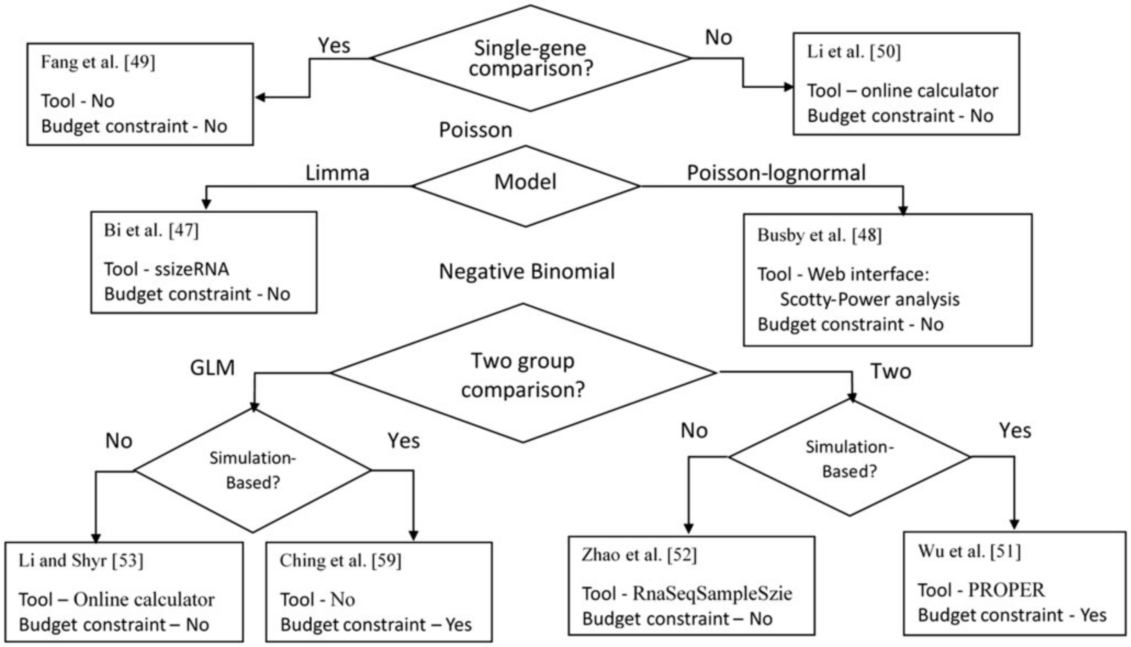

Figure 1. Flow chart for selecting the proper power/sample size (PS) method for RNA-seq experiment. The researcher first needs to decide which statistical model will be used to describe the count data. If the limma model is selected, the method proposed by Bi *et al.* [14] is the correct tool for conducting a PS analysis; if Poisson-lognormal model is selected, the method proposed by Busy *et al.* [16] is the correct tool for conducting a PS analysis; if Poisson model is selected for conducting single-gene comparison, the model proposed by Fang *et al.* [17] is the correct tool; if Poisson model is selected for conducting multiple-gene comparisons, the model proposed by Li *et al.* [18] is the correct tool. If the negative binomial model is selected to conduct two-group comparisons, the method proposed by Wu *et al.* [19] is the simulation-based tool; if the negative binomial model is selected to conduct two-group comparison, the method proposed by Zhao *et al.* [20] is the correct tool; if the negative binomial model is selected to conduct multiple-group comparisons, the method proposed by Ching *et al.* [19] is the simulation-based tool; and if the negative binomial model is selected to conduct multiple-group comparison, the method proposed by Li and Shyr [21] is the correct tool.

dependent on several parameters: the depth of sequencing of the normal/reference sample, the depth of the tumor sample, the expected tumor contamination rate in the normal sample and the expected mutation percentage in the sample.

The power to detect somatic mutations can be modeled based on Fisher's exact test for two proportions. The power analysis using Fisher's exact test is rather complicated (see Sahai and Khurshid [30] for details). Fortunately, the power analysis for Fisher's exact test has been implemented in the R package exact2×2 and the SAS procedure PROC POWER with option TWOSAMPLEFREQ.

After one successfully identifies SNPs or somatic mutations, one can estimate the population frequency of these single-nucleotide variants (SNVs). The power to correctly estimate population SNV frequency is dependent on the sample size and the true population SNV frequency. Moreover, the power is also affected by the accuracy of the variant callers, and thus, it depends on the exact methodology applied. Assuming the accuracy among the variant callers is similar, the power analysis in this scenario becomes a traditional statistical problem of sample size needed for estimating a proportion. This can be implemented in the R package pwr and the SAS procedure PROC POWER with option ONESAMPLEFREQ.

Power to detect association for common variants

A parameter considered by some power calculators is budget, which is not directly considered during power computation in GWAS. For HTS, under a fixed budget, the investigator can choose to either sequence more samples at a lower depth or to sequence fewer samples at a higher depth, a consideration that is not applicable for genotyping arrays. Increased depth will increase the power of detecting a variant in a sample, and increased sample size will increase the power to correctly identify the variant allele frequency in the population. This trade-off has been thoroughly discussed by Shen *et al.* [31]. Based on their model of power analysis, the authors found that the maximum power for detecting phenotype association can be obtained by selecting the optimal balance between the average depth and the number of samples. A practical approach is to only sequence cases, and use existing public data, such as the 1000 Genome Project [32], as the control group. While this may work for phenotypes with a low population frequency, where the control group is nearly identical to the general population, this approach fails for common phenotypes. Furthermore, it is highly susceptible to artifacts because of differences between the local study population and the reference population. When both cases and controls have to be sequenced and the cost is proportional to the number of subjects, the optimal fraction of cases with the maximum power of detecting associations has been suggested to be 1/*e*, where *e* is the base of the natural logarithm [31].

Power to detect association for rare variants (aggregated power)

Traditional GWAS aims to identify common variants in common diseases using genotyping arrays under the common disease–common variant (CDCV) hypothesis, which states that common diseases are caused by common variants. Under the CDCV hypothesis, each common variant has a small-to-modest additive or multiplicative effect on disease phenotype [33, 34]. An alternative hypothesis is the common disease–rare variant hypothesis, which states that risks for common diseases may be caused by multiple rare variants in the same gene or same pathway [35]. Furthermore, rare Mendelian diseases are usually caused by rare variants with large effects.

Genotyping arrays are not designed to detect all common variants (even after imputation with linkage disequilibrium) or the rare variants with a large effect [36]. Even genotyping arrays that include rare variants in their design can only detect a small fraction of the true rare variants in a study population. HTS, on the other hand, can interrogate the entire genome or exome at the single-nucleotide resolution. However, because of a high per sample unit price, using sequencing to perform a large GWAS is still impractical economically. HTS (either whole genome or exome) has been used to study rare diseases or common diseases with limited sample sizes, which limit the power of a traditional univariate regression analysis [37]. To overcome the limitation in sample size in HTS data, a wide range of aggregated methods has been developed, such as CAST [38], Combined and Multivariate Collapsing [37], weighted sum method [39], variable threshold [40], rare variant, weighted aggregate statistic [41], kernel-based adaptive clustering method [42], C-alpha [43], data-adaptive sum test [44], RareCover test [45], replication-based test [46] and SNP-set Kernel Association Test (SKAT) [47].

The power increase in an aggregated approach can be attributed to two reasons. First, by collapsing SNPs from a genomic region of interest, usually defined as a gene or pathway, into one score, the number of tests performed is substantially reduced, thus alleviating the burden of multiple testing corrections. Second, it is assumed that rare variants with different genomic positions in a gene may disrupt the function of the gene, as this has often been observed in Mendelian diseases such as cystic fibrosis [48]. Testing at the single-variant level will not capture the collective effect of these variants at the gene level, and collapsing these SNPs to one value will increase the signal strength. Power analysis for an aggregated test is complex, and restricted to many assumptions. Within a region of interest, the effect of variants may be nonuniform, or it could even be of the opposite direction (detrimental versus protective) and noncausal. Furthermore, many of the aggregated methods use intractable mathematical formulas or calculations, making the power analysis difficult and impractical.

Currently, there are several available approaches for computing power with aggregated approaches. Lee *et al.* [49] derived analytical formulas to compute power for SKAT analysis based on an approximate noncentral chi-square distribution under distinct scenarios: retrospective case-control studies, rare variant studies and average power across different regions. This method is implemented in the SKAT R package. However, Wu *et al.* [50] showed that the power based on the analytical approach proposed by Lee *et al.* (2012) could be inflated when the significance level is small. To accurately and efficiently compute power, Wu *et al.* proposed an exact method based on a new noncentral chi-square approximation. To accurately calculate the power for SKAT, Wu *et al.*'s method is more appropriate. The implementation of Wu *et al.*'s method is available as an R package KATSP. An alternative to the analytical approach is the simulation approach. SPS [51] is a Monte Carlo simulation-based power analysis designed for SKAT with an advanced graphical user interface. Moreover, SPS also can be used to estimate the power for meta-analysis. Wang *et al.* developed SEQPower [52] that can perform power analysis for allele frequency and quantitative trait-based aggregated tests using a Monte Carlo approach applying forward-time simulated [53] sequencing data. SPS and SEQPower are recommended when the aggregated test is not focusing on SKAT.

RNA

RNA-seq uses the HTS technology to sequence complementary DNA reverse transcribed from RNA. The raw data of RNA-seq contain millions of short reads, which are aligned back to a reference genome or transcriptome. The reads aligned to each gene serve as measurements of the mRNA expression levels. Several power analysis methods have been proposed. Each method has its own advantage and limitation. Selecting an appropriate power assessment method is crucial to the study design. To provide researchers with better guidance for selecting the tools to conduct a power analysis, we produced a flow chart, shown in Figure 1. The researcher first needs to decide which statistical model will be used to describe the count data. If the Limma model is selected, the method proposed by Bi *et al.* [15] is the correct tool for conducting a PS analysis; if Poisson-lognormal model is selected, the method proposed by Busy *et al.* [16] is the correct tool for conducting a PS analysis; if Poisson model is selected for conducting a single-gene comparison, the model proposed by Fang *et al.* [17] is the correct tool; if the Poisson model is selected for conducting a multiple-gene comparison, the model proposed by Li *et al.* [18] is the correct tool; if the negative binomial model is selected to conduct a two-group comparison, the method proposed by Wu *et al.* [19] is the appropriate simulation-based tool; if the negative binomial model is selected to conduct a two-group comparison, the method proposed by Li *et al.* [20] is the correct tool; if the negative binomial model is selected to conduct a multiple-group comparison, the method proposed by Ching *et al.* [19] is the simulation-based tool; and finally, if the negative binomial model is selected to conduct a multiple-group comparison, the method proposed by Li and Shyr [21] is the correct tool.

Poisson model

In statistics, the Poisson distribution is widely used to model counting processes. Because RNA-seq data can be represented as read counts, Fang *et al.* [17] used Poisson distribution to model count data and derived a sample size formula based on a Wald test or a likelihood ratio test (LRT) for single-gene differential expression analysis. There are two limitations for this method. In reality, in RNA-seq data analysis, tens of thousands of genes are examined and tested simultaneously. Thus, the correction for multiple testing needs to be considered. For multiple gene comparison, Li *et al.* [18] derived sample size calculation formulas based on the most common test statistics, including the Wald test and Rao's score test, log transformation of score test and log transformation of Wald test. Moreover, because it is difficult to derive a closed form to calculate the sample size based on a LRT, Li *et al.* [18] proposed a numerical approach to address this issue. Their method was implemented as an online calculator, RNAseqPS [54]. Currently, those are the only two available methods that can assess power for tests of differential expression from RNA-seq data based on a Poisson model.

Negative binomial model

It has been repeatedly shown that RNA-seq data exhibit an overdispersed read count distribution [55, 56], which means that the variance of sequence counts exceeds the mean. The power analysis methods based on a Poisson distribution are unable to take this variability into account. To compensate for this overdispersion, the negative binomial distribution is a more flexible for describing the mean–variance relationship. Based on a negative binomial distribution, Hart *et al.* [57] proposed a power analysis method based on the score test for single-gene differential expression analysis. This method has been implemented in Bioconductor as part of the RNASeqPower package. To handle multiple gene comparisons, Li *et al.* [58] proposed a power analysis method, while also controlling for the false discovery rate based on the exact test implemented in Bioconductor package edgeR [55]. However, because the individual power analysis for the exact test involved infinite sums, and the overall power of the study is estimated by summing the individual power, Li *et al.*'s method is computationally expensive. Thus, in the same publication, to alleviate the computational burden, Li *et al.* [58] further proposed a method for calculating a conservative sample size based on the minimum average read counts in the control group, the minimum fold change and the maximum dispersion. Instead of using a single value for the maximum dispersion and the minimum average read counts, Zhao *et al.* [20] implemented Li *et al.*'s method to develop an algorithm based on the distributions of read counts and dispersion estimated from prior data. This method is implemented in the Bioconductor package RnaSeqSampleSize.

Similar to DNA sequencing studies, budget plays a significant role in the design of the RNA-seq study. To incorporate budget as part of power analysis, Wu *et al.* [19] introduced the concepts of stratified power by coverage or biological variation and cost of false discovery, then proposed a simulation-based method for power analysis. The method was implemented as a Bioconductor package, PROPER [19].

The aforementioned methods based on the negative binomial model are designed for assessing the differential expression between two groups. For complex RNA-seq experimental design involving multiple group comparisons, Ching *et al.* [59] used a simulation-based method under a generalized linear model framework to perform power analysis for a given budget constraint. For power analysis, Wu *et al.* [19] and Ching *et al.* [59] considered a wide range of differential expression analysis packages including DEseq [60], edgeR [55], DSS [61], DESeq2 [62], EBSeq [63] and SSeq [64], respectively. Thus, those methods offered great flexibility in downstream analysis. Most recently, to avoid complex mathematical approximations, Li and Shyr [21] proposed a power analysis method using an LRT under the generalized linear model. Because the Bioconductor packages edgeR, DESeq and DEseq2 provided statistical methods using an LRT for assessing the differential expression analysis, this method is directly applicable. This method was implemented in a Web-based user interface (http://140.116.152.140/shiny/App/GLM/).

Poisson-lognormal distribution model

Busby *et al.* [16] observed that in data sets, the distribution of the log read counts appears to be approximated by a truncated normal distribution. Thus, it is reasonable to model gene expression as a lognormal distribution. However, the abundances of gene expression are measured with read counts. To combine those observations, Busby *et al.* [16] assumed that the read count follows a Poisson distribution and that the gene expression follows a lognormal distribution. Thus, the distribution of read counts is more appropriately modeled by a Poisson-lognormal distribution. A sample size calculation formula based on a *t*-test for assessing a single-gene differential expression between two groups was derived as following:

$$\mathrm{T_v}\left(\mathrm{t}_{\alpha/2,\upsilon}|\frac{\delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\right)-\mathrm{T_v}\left(-\mathrm{t}_{\alpha/2,\upsilon}|\frac{\delta}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\right),$$

where δ is the effect size, $\mathrm{T_v}(.|\theta)$ is distribution function of the noncentral *t*-distribution with noncentral parameter θ and degrees of freedom v, $\mathrm{t}_{\alpha/2}$ are 100(1-α/2) percentile of the *t*-distribution with v degrees of freedom. To calculate the total power for multiple gene comparison, Busby *et al.* [16] proposed that the overall power of the study is the mean power for assessing a single-gene differential expression. Their method is implemented in a Web interface: Scotty-Power analysis for RNA-seq experiments.

Limma model

Limma is a linear model-based method originally designed for microarray gene expression analysis [65]. The defining feature of the limma approach is the utilization of an empirical Bayes method for borrowing information across genes, making the analyses stable even for experiments with a small number of arrays [66]. Recently, Ritchie *et al.* expanded the function of the Bioconductor package limma to analyze RNA-seq data [67].

Based on the limma approach, Law *et al.* [68] introduced a voom method, which applies precision weights to account for the mean–variance relationship of the log count data. Based on the voom and limma approaches, Bi *et al.* [15] proposed a one-time simulation method for power analysis to assess the differential expression between two groups. Because of the flexibility of the simulation approach, this method can be extended to other experimental designs, such as paired-sample or multiple treatment comparisons. This method has been implemented in the R package ssizeRNA. In the same study, Bi *et al.* conduct a simulation experiment to compare the performances among the R packages ssizeRNA, RnaSeqSampleSize and PROPER. Bi *et al.* showed that ssizeRNA provided a more accurate estimate of power/sample size than RnaSeqSampleSize; ssizeRNA and RnaSeqSampleSize provided results much faster than PROPER. Thus, Bi *et al.* concluded that ssizeRNA works best when both accuracy and computation time are considered.

Microbiome

Another popular application of HTS is to study the microbiome, known as microbiome sequencing. Microbes (bacteria, fungi and archaea) can be found throughout the human body. Increasing interest in these microbes' contribution to disease has propelled the development of a series of microbiome sequencing analysis pipelines [69, 70]. The basic goals of the microbiome studies are to identify microbiome species diversity within one sample (α-diversity), or multiple samples (β-diversity) or the relative abundance of one or multiple microbes between two groups. Microbiome data are commonly modeled with a multinomial distribution because the abundance of each microbe is represented as a percentage of the total reads, and the sum of the total microbe abundance within a sample is 1. The probability mass function of the multinomial distribution for microbiome data can be described as follows:

$$\mathrm{f}(\mathrm{x}_1,\ldots,\mathrm{x}_J;\pi_1,\ldots,\pi_J)=\frac{\mathrm{N}!}{\Pi_{i=1}^J\mathrm{x}_i!}\Pi_{i=1}^J\pi_i^{\mathrm{x}_i},$$

where J is the number of taxa, $\mathrm{N}=\sum_{i=1}^J \mathrm{x}_i$ is the total taxa count and $(\pi_1,\ldots,\pi_J)$ are the abundance of species with $\sum{}_{i=1}^J\pi_i=1$.

Power to detect α-diversity

Thompson (1987) [71] provided a procedure for estimating the parameters of a multinomial distribution. Because the

α-diversity is a function of the species' proportional abundance, Thompson's method can be applied directly. This method can used to identify the number of detectable microbes within one sample. In this scenario, the total read count is the determining factor. Similar to RNA-seq experiments, the total number of reads sequenced directly affects the ability to sequence the low-abundant genes or microbes. The power for the measurement of α-diversity can be defined as the proportion of abundance.

Power to detect β-diversity

Multiple distance-based methods [72–75] have been proposed for estimating β-diversity. Details regarding those distance-based methods can be found in reference [76]. Distance-based methods have two limitations: they are underpowered when the single distance is poorly chosen and they cannot handle the variables that correlate with both the covariates of interest and the microbiome composition [76]. To address these issues, Tang *et al.* [76] proposed a new distance-based method to test the association of microbial communities based on the permutation multivariate analysis of variance (PERMANOVA). Their method is implemented in software, which is available at https://medschool.vanderbilt.edu/tang-lab/software/miProfile. With the rapid growth of estimating β-diversity, the power analysis has lagged behind. Kelly *et al.* proposed a method for generating the pairwise subject-to-subject distance matrix that permits modeling within-group distance according to prespecified parameters. Based on the simulated distance matrix, the power of PERMANOVA can be calculated for a given group-level effect size, which is quantified by the adjusted coefficient of determination, Omega squared. Their method is implemented in the R package micropower [77].

Power to detect relative abundance

In addition to α-diversity and β-diversity, the relative abundance of a single or multiple microbes between two or multiple groups can also be interesting to examine [78, 79]. If we do not assume a Dirichlet distribution of the microbiome data, the microbes can be tested for differential abundance individually. In such a situation, the focus of interest is the differential abundance of a single microbe. This problem could be simplified to the same scenario as the differential gene expression analysis in RNA-seq. Thus, we can apply power/sample analyses designed for RNA-seq data. However, Rosa *et al.* [80] showed that microbiome data are better modeled with a Dirichlet multinomial distribution when the overdispersion is present. Rosa *et al.* [80] proposed a power method based on a Dirichlet multinomial distribution, which was implemented in the R package HMP.

Chromatin immunoprecipitation sequencing

ChIP-seq experiments use chromatin immunoprecipitation (ChIP) with HTS to identify the binding sites of DNA-associated proteins. ChIP-seq data is similar to RNA-seq data, involving quantifying the data as read count per peak instead of per gene. A peak in this context is a genomic region that has been enriched with aligned reads as evidence of a DNA-binding protein in that region. ChIP-seq data has been modeled using a local Poisson model [81, 82]. Zuo *et al.* [83] developed a statistical framework for ChIP-seq experiments based on the assumption that the reads are generated by local Poisson processes with shared Gamma prior distributions. To control the false discovery rate, they defined a conditional power function and proposed a numerical algorithm to compute the following conditional posterior power.

$$\frac{\sum_{i=1}^{n} w_i E[p\{Y_i > T_i(\alpha_q, \gamma, \tau)|\lambda_i^y > re_0\mu_i \vee \tau, B_i \neq 0\}|Y_i = y_i]}{\sum_i^n w_i}$$

where $w_i = p\{\lambda_i^y > re_0\mu_i \vee \tau,\ B_i \neq 0|Y_i = y_i\}$, α_q the significance level, Y_i is the observed ChIP counts, γ is the fold change, τ is a minimum intensity, e_0 is a normalizing factor reflecting the proportion of background reads, $x \vee y = \max(x, y)$, B_i is used to indicate the enrichment state of bin *i*. This method was implemented in the R package CSSP. In addition, to measure the reproducibility of the finding from an experiment design, Li *et al.* [84] proposed a reproducibility score, the irreproducible discovery rate (IDR). IDR could be considered as a post-sequencing evaluation of the power a ChIP-Seq analysis.

Discussion

HTS technology has undoubtedly reshaped the landscape of genomics. The true advantage of HTS lies in its versatility, and the way it allows itself to be adapted for a wide range of applications. Each type of application of HTS aims to examine a unique

Table 1. Power/sample size computation tools

Article	Test statistics	Design	Test	Sequencing	Package/software
[15]	Binomial test	Case-control	Association	DNA	OPERA
[49]	Score test (SKAT)	Case-control	Aggregated test	DNA	None
[52]	Simulation	MCMC	Aggregated test	DNA	SEQPower
[41]	Simulation	MCMC	Aggregated test	DNA	SPS
[50]	Score test (SKAT)	Linear model	Aggregated test	DNA	R package KATSP
[17]	Wald and LRT	Case-control	Single gene	RNA	None
[18]	Wald, Score and LRT	Case-control	Multigene	RNA	None
[57]	Score	Case-control	Single gene	RNA	RNASeqPower
[20]	Exact	Case-control	Multigene	RNA	RnaSeqSampleSize
[19]	Simulation	Case-control	Multigene	RNA	PROPER
[59]	Simulation	Linear model	Multigene	RNA	powerSampleSizeCalculator
[16]	*t*-test	Case-control	Multigene	RNA	Scotty-Power analysis
[15]	Voom	Linear model	Multi gene	RNA	ssizeRNA
[77]	PERMANOVA	Multiple groups	β-diversity	Microbiome	R package micropower
[80]	Wald test	Dirichlet	Relative abundance	Microbiome	R package HMP
[83]	Exact text	Gamma	Peak difference	ChIP-seq	R package CSSP

biological problem. Combined with the complex format of HTS data, and the distinct statistical assumption behind the different types of HTS data analysis, the power analysis for HTS-based experiments has been a challenging problem. In this review, we have described the power analysis methods for four types of HTS applications: DNA sequencing, RNA-seq, microbiome sequencing, and ChIP-seq. Each of these HTS application can be further divided into subcategories depending on the goals of the experiments. In each case, the definition of power can vary greatly based on the goal of the study. The existing tools developed for power analysis have been listed in Table 1.

For complex study design and tests, as demonstrated in the DNA and RNA sequencing power analyses, simulation-based methods are the only feasible approach to model intractable mathematical computation. The major downside for simulation-based approaches is the long time required for their accurate calculation. Furthermore, there are many more less well-known applications of HTS technology whose power analyses have not been properly studied, such as GRO-seq, CLIP-seq, etc. As demand for these types of HTS applications increases, and new applications are created, additional power computation methods for these types of methods will need to be developed.

Key Points

- The power analysis for HTS -based experiments is more complicated than for microarray based experiments because of extra parameters such as read depth, read dispersion, etc.
- The power analysis varies greatly based on the type of the HTS data and the goal of the experiment.
- The power analysis is an essential part of an experiment's success.

Funding

This study was partially supported by Cancer Center Support Grant (P30 CA068485) from National Cancer Institute.

References

1. Ng SB, Turner EH, Robertson PD, *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;**461**:272. U153.
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
3. Li Y, Tollefsbol TO. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol* 2011;**791**:11–21.
4. Di Bella JM, Bao Y, Gloor GB, *et al*. High throughput sequencing methods and analysis for microbiome research. *J Microbiol Methods* 2013;**95**:401–14.
5. Danko CG, Hyland SL, Core LJ, *et al*. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* 2015;**12**:433.
6. Jothi R, Cuddapah S, Barski A, *et al*. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.
7. Hafner M, Landthaler M, Burger L, *et al*. Transcriptome-wide identification of RNA-binding protein and MicroRNA target sites by PAR-CLIP. *Cell* 2010;**141**:129–41.
8. Klein RJ. Power analysis for genome-wide association studies. *BMC Genet* 2007;**8**:58.
9. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. *Genomics Inform* 2012;**10**:117–22.
10. Spencer CC, Su Z, Donnelly P, *et al*. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet* 2009;**5**:e1000477.
11. Skol AD, Scott LJ, Abecasis GR, *et al*. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies (vol 38, pg 209, 2006). *Nat Genet* 2006;**38**:390.
12. Feng S, Wang SC, Chen CC, *et al*. GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. *BMC Genet* 2011;**12**:12.
13. Visscher PM, Hemani G, Vinkhuyzen AAE, *et al*. Statistical power to detect genetic (Co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* 2014;**10**:e1004269.
14. Gauderman W, Morrison J. QUANTO 1.1: a computer program for power and sample size calculations for genetic-epidemiology studies. http://hydra.usc.edu/gxe. 2006.
15. Bi R, Liu P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* 2016;**17**:146.
16. Busby MA, Stewart C, Miller CA, *et al*. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* 2013;**29**:656–7.
17. Fang Z, Cui X. Design and validation issues in RNA-seq experiments. *Brief Bioinform* 2011;**12**:280–7.
18. Li C-I, Su P-F, Guo Y, *et al*. Sample size calculation for differential expression analysis of RNA-seq data under Poisson distribution. *Int J Comput Biol Drug Design* 2013;**6**:358–75.
19. Wu H, Wang C, Wu ZJ. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* 2015;**31**:233–41.
20. Zhao S, Guo Y, Sheng Q, Shyr Y. RNASeqSampleSize. R package version 1.4.2. https://cqs.mc.vanderbilt.edu/shiny/RNAseqPS/.
21. Li CI, Shyr Y. Sample size calculation based on generalized linear models for differential expression analysis in RNA-seq data. *Stat Appl Genet Mol Biol* 2016;**15**:491–505.
22. Chow S-C, Wang H, Shao J. *Sample size Calculations in Clinical Research*. CRC press, 2007. https://www.crcpress.com/Sample-Size-Calculations-in-Clinical-Research-Second-Edition/Chow-Wang-Shao/p/book/9781584889823.
23. Guo Y, Samuels DC, Li J, *et al*. Evaluation of allele frequency estimation using pooled sequencing data simulation. *ScientificWorldJournal* 2013;**2013**:895496.
24. McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.
25. Gustin JP, Cosgrove DP, Park BH. The PIK3CA gene as a mutated target for cancer therapy. *Curr Cancer Drug Targets* 2008;**8**:733–40.
26. Cheng S, Chu P, Hinshaw M, *et al*. Frequency of mutations associated with targeted therapy in malignant melanoma patients. *J Clin Oncol* 2011;**29**.
27. Sheng Q, Zhao S, Li CI, *et al*. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics* 2016;**107**:163–9.
28. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun* 2015;**6**:8971.
29. Su XP, Zhang L, Zhang JP, *et al*. PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* 2012;**28**:2265–6.
30. Sahai H, Khurshid A. Formulae and tables for the determination of sample sizes and power in clinical trials for testing

differences in proportions for the two-sample design: a review. *Stat Med* 1996;**15**:1–21.

31. Shen Y, Song R, Pe'er I. Coverage tradeoffs and power estimation in the design of whole-genome sequencing experiments for detecting association. *Bioinformatics* 2011;**27**:1995–7.
32. Abecasis GR, Auton A, Brooks LD, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
33. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001;**17**:502–10.
34. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;**6**:95–108.
35. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;**11**:415–25.
36. Gibson G. Rare and common variants: twenty arguments. *Nat Rev Genet* 2012;**13**:135–45.
37. Li BS, Leal SM. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet* 2008;**83**:311–21.
38. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat Res Fundam Mol Mech Mutagen* 2007;**615**:28–56.
39. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;**5**:e1000384.
40. Price AL, Kryukov GV, de Bakker PIW, *et al*. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010;**86**:832–8.
41. Sul JH, Han B, He D, *et al*. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics* 2011;**188**:181. U298.
42. Liu DJJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 2010;**6**:e1001156.
43. Neale BM, Rivas MA, Voight BF, *et al*. Testing for an unusual distribution of rare variants. *PLoS Genet* 2011;**7**:e1001322.
44. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;**70**:42–54.
45. Bhatia G, Bansal V, Harismendy O, *et al*. A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput Biol* 2010;**6**:e1000954.
46. Ionita-Laza I, Buxbaum JD, Laird NM, *et al*. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet* 2011;**7**:e1001289.
47. Ionita-Laza I, Lee S, Makarov V, *et al*. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet* 2013;**92**:841–53.
48. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* 2015;**7**:16.
49. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;**13**:762–75.
50. Wu B, Pankow JS. On sample size and power calculation for variant set-based association Tests. *Ann Hum Genet* 2016;**80**:136–43.
51. Li J, Sham PC, Song YQ, *et al*. SPS: a simulation tool for calculating power of set-based genetic association tests. *Genet Epidemiol* 2015;**39**:395–7.
52. Wang GT, Li B, Santos-Cortez RPL, *et al*. Power analysis and sample size estimation for sequence-based association studies. *Bioinformatics* 2014;**30**:2377–8.
53. Peng B, Liu XM. Simulating sequences of the human genome with rare variants. *Hum Hered* 2010;**70**:287–91.
54. Guo Y, Zhao S, Li CI, *et al*. RNAseqPS: a web tool for estimating sample size and power for RNAseq experiment. *Cancer Inform* 2014;**13**:1–5.
55. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
56. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 2007;**23**:2881–7.
57. Hart SN, Therneau TM, Zhang YJ, *et al*. Calculating sample size estimates for RNA sequencing data. *J Comput Biol* 2013;**20**:970–8.
58. Guo Y, Li J, Li CI, *et al*. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* 2013;**29**:1210–11.
59. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* 2014;**20**:1684–96.
60. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
61. Wu H, Wang C, Wu ZJ. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* 2013;**14**:232–43.
62. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
63. Leng N, Dawson JA, Thomson JA, *et al*. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments (vol 29, pg 1035, 2013). *Bioinformatics* 2013;**29**: 2073.
64. Yu DN, Huber W, Vitek O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* 2013; **29**:1275–82.
65. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 2004;**3**:Article3.
66. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* 2005;**21**:2067–75.
67. Ritchie ME, Phipson B, Wu D, *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
68. Law CW, Chen Y, Shi W, *et al*. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29.
69. Schloss PD, Westcott SL, Ryabin T, *et al*. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;**75**:7537–41.
70. Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**:335–6.
71. Thompson SK. Sample-size for estimating multinomial proportions. *Am Stat* 1987;**41**:42–6.
72. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 2005;**71**:8228–35.
73. Lozupone CA, Hamady M, Kelley ST, *et al*. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl Environ Microbiol* 2007;**73**:1576–85.

74. Chen J, Bittinger K, Charlson ES, *et al*. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 2012;**28**:2106–13.
75. Evans SN, Matsen FA. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *J R Stat Soc Ser B Stat Methodol* 2012;**74**:569–92.
76. Tang ZZ, Chen G, Alekseyenko AV. PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics* 2016;**32**:2618–25.
77. Kelly BJ, Gross R, Bittinger K, *et al*. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics* 2015;**31**:2461–8.
78. Chen J, Li HZ. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat* 2013;**7**:418–42.
79. Wang Y, Naumann U, Wright ST, *et al*. mvabund- an R package for model-based analysis of multivariate abundance data. *Methods Ecol Evol* 2012;**3**:471–4.
80. La Rosa PS, Brooks JP, Deych E, *et al*. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One* 2012;**7**:e0052078.
81. Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biology* 2008;**9**:R137.
82. Ji H, Jiang H, Ma W, *et al*. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 2008;**26**:1293–300.
83. Zuo C, Keles S. A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* 2014;**30**:753–60.
84. Li QH, Brown JB, Huang HY, *et al*. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 2011;**5**: 1752–79.