Bioinformatics, 35(3), 2019, 529–531
doi: 10.1093/bioinformatics/bty542
Advance Access Publication Date: 19 July 2018
Applications Note

OXFORD

Genetics and population analysis

MendelProb: probability and sample size calculations for Mendelian studies of exome and whole genome sequence data

Zongxiao He[1], Lu Wang[2], Andrew T. DeWan[3] and Suzanne M. Leal[1,*]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Center for Statistical Genetics, Houston, TX 77030, USA, [2]Department of Statistics, Rice University, Houston, TX 77005, USA and [3]Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT 06510, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on March 26, 2018; revised on June 1, 2018; editorial decision on June 26, 2018; accepted on July 18, 2018

Abstract

Motivation: For the design of genetic studies, it is necessary to perform power calculations. Although for Mendelian traits the power of detecting linkage for pedigree(s) can be determined, it is also of great interest to determine the probability of identifying multiple pedigrees or unrelated cases with variants in the same gene. For many diseases, due to extreme locus heterogeneity this probability can be small. If only one family is observed segregating a variant classified as likely pathogenic or of unknown significance, the gene cannot be implicated in disease etiology. The probability of identifying several disease families or cases is dependent on the gene-specific disease prevalence and the sample size. The observation of multiple disease families or cases with variants in the same gene as well as evidence of pathogenicity from other sources, e.g. *in silico* prediction, expression and functional studies, can aid in implicating a gene in disease etiology. MendelProb can determine the probability of detecting a minimum number of families or cases with variants in the same gene. It can also calculate the probability of detecting genes with variants in different data types, e.g. identifying a variant in at least one family that can establish linkage and more the two additional families regardless of their size. Additionally, for a specified probability MendelProb can determine the number of probands which need to be screened to detect a minimum number of individuals with variants within the same gene.

Results: A single Mendelian disease family is not sufficient to implicate a gene in disease etiology. It is necessary to observe multiple families or cases with potentially pathogenic variants in the same gene. MendelProb, an R library, was developed to determine the probability of observing multiple families and cases with variants within a gene and to also establish the numbers of probands to screen to detect multiple observations of variants within a gene.

Availability and implementation: https://github.com/statgenetics/mendelprob

Contact: sleal@bcm.edu

1 Introduction

For studies of Mendelian diseases, pathogenic or potentially pathogenic variant identification is often performed using exome and whole genome sequence data generated using DNA samples from families with multiple affected members, trios or single cases. Filtering approaches are used to analyze exome and whole genome sequence data where variants are selected based on (i) having very low minor allele frequencies (MAFs), e.g. <0.005 in every ancestry group in gnomAD; (ii) bioinformatic tool predictions; (iii) mode of inheritance, e.g. for autosomal recessive (AR) diseases either

compound heterozygous or homozygous; (iv) being *de novo* and (v) if multiple affected family members are sequenced, the variants should be shared amongst family members. For families, segregation of identified variant(s) is evaluated by sequencing all available informative family members. Linkage analysis should be used for pedigrees to evaluate the statistical significance of the identified variant(s) and to aid in their classification. See Ott *et al.* (2015), for a more complete overview of filtering and linkage analysis.

Although there are criteria to classify variants as benign, likely benign, variant of unknown significance (VUS), likely pathogenic or pathogenic, there is no precise classification system to implicate a gene in disease etiology (Richards *et al.*, 2015). If at least one variant within a gene has been identified that can be classified as pathogenic, the gene can be implicated in disease etiology. However, if none of the variants within a gene are classified as pathogenic, there are no strict criteria to implicate a gene in disease etiology, i.e. necessary number of observations of VUS or likely pathogenic variants.

It is necessary to establish that genetic studies are sufficiently powered. For association studies, power is estimated for a specified significance level, genetic model, disease prevalence and disease and variant MAFs. For pedigree-based Mendelian diseases, gene dropping simulations studies can be used to empirically estimate power for a specified mode of inheritance, penetrance model and disease and variant MAFs (Boehnke, 1986). For Mendelian disease studies, even if a single pedigree can be observed which meets the significant LOD score threshold criterion i.e. 3.3 (Lander *et al.*, 1995), this evidence is not sufficient to demonstrate a gene's involvement in disease etiology (Ott *et al.*, 2015). It is necessary to observe multiple disease families or cases with variants in the same gene and have additional supporting evidence from bioinformatic tools, expression and functional studies (Richards *et al.*, 2015). For many diseases, e.g. intellectual disabilities, non-syndromic hearing impairment (HI), retinitis pigmentosa, there is a high degree of locus heterogeneity making it challenging to observe multiple families or cases with pathogenic or potentially pathogenic variants in the same gene.

2 Description

MendelProb can be used to determine the probability of identifying at least N families or unrelated cases with variants in the same gene, or it can be specified that at least X of these families must be of a specific type, e.g. sufficiently large to produce a significant LOD score of 3.3. If a family with potentially pathogenic variant(s) has already been identified, for a specific probability MendelProb can determine how many additional probands should be sequenced to observe variant(s) in the same gene, or it can be determined how many probands need to be screened to identify $\geq N$ individuals with variants in the same gene. The probands can be either an affected family member or a case. Even if sequence data are generated on more than one family member to perform filtering, when calculating probabilities or number of subjects to be sequenced, each family is counted only once and neither size nor structure of the family will impact the results.

As an example, a study on non-syndromic HI (NSHI) is used where a total of 125 families and 500 cases of African-American ancestry are ascertained ($N = 625$). The probability of detecting at least four probands with potentially pathogenic variant(s) in the same gene that explain 0.5% of NSHI among the 625 probands sequenced is 17.5%. If the criterion is relaxed and it is only necessary to detect at least two probands with variant(s) in the same gene the probability is 82%. Using this relaxed criterion, if there are 100 NSHI genes each explaining 0.5% of NSHI, then potentially 82 genes could be identified that have at least two probands with variants in the same gene. A recent study observed that for 74% of African-Americans, HI was not due to a known gene (Sloan-Heggen *et al.*, 2016). Therefore, of the 82 genes, 60 novel NSHI genes could potentially be identified. Additionally, for this study for a gene that explains 1% of NSHI, the probability is 71% to detect at least two probands with variants in the same gene with at least one proband being from one of the 125 families. If the gene only explains 0.5% of NSHI the probability drops to 44%.

The number of probands that need to be screened to detect at least N probands with variants in the same gene was also determined. If it is desired to identify at least two probands with variant(s) in the same gene which underlies 1% of NSHI with a probability of 80%, a total of 298 probands would need to be sequenced. It can also be determined how many probands need to be sequenced to find at least N additional probands with variant(s) in the same gene, if a family has already been identified. These calculations can also be performed if it is desired to detect at least N probands with variants in the same gene with at least X being of a specific type. For this scenario, the proportion of probands in each category must be specified. If a sample is screened with one-third families and two-third cases to detect three probands with variant(s) in a gene which is responsible for 0.5% of NSHI with 80% probability with at least one proband being from a family, 1108 probands must be sequenced. If the proportions are changed to one-half families and one-half cases, then 923 probands need to be screened.

3 Discussion

For the previous calculations, the percent of disease caused by a gene is used to determine probabilities and sample sizes, neither mode of inheritance, variant frequencies nor disease penetrance need to be specified, since the percent of disease due to a gene is dependent on these parameters. When calculating probabilities and sample sizes, it is advisable to use a range of low frequencies for the percent of disease caused by a single gene, e.g. 0.25–2.0%. For a study, it is possible to discover multiple genes each explaining a different proportion of disease etiology for which the contribution of each gene will be unknown *a priori*.

It is also possible to estimate the probability of identifying a variant underlying autosomal dominant (AD) or AR diseases in multiple families or cases or determine the number of probands which need to be sequenced. For these calculations, p^2, where p represents the MAF, is used for a fully penetrant AR disease and $(1-q^2)$ for a fully penetrant AD disease. For an AR variant with reduced penetrance, $F_{DD}(p^2)$ is used for the frequency in the calculations, where F_{DD} is the proportion of individuals who are homozygous for the variant of interest who develop disease. For AD variants with reduced penetrance, when penetrance is equivalent for the homozygous and heterozygous state $F_{D*}(1-q^2)$ is used for frequency calculations, where F_{D*} is the proportion of homozygous or heterozygous variant individuals who develop disease. A caveat of performing these calculations is the assumption that Hardy–Weinberg Equilibrium cannot be violated and the probands which are selected for sequencing must be obtained from families or cases which have disease etiology that is either AR or AD.

If probands are being screened, usually it is of interest to find N probands and not $\geq N$. To identify exactly N probands, sequencing must be done sequentially. For sequencing studies this is not practical, therefore sample size estimates for a fixed N are not performed.

MendelProb can be used for grant proposals and designing studies for Mendelian traits to determine the probability of being able to identify multiple probands or determine the sample size that should be screened to detect several probands with pathogenic or potentially pathogenic variants in the same gene. These estimates can be used on their own or in conjunction with power calculations.

Funding

This work is supported by National Institute of Health [grants numbers R01 AG058131, R01 DC011651, R01 DC003594, R01 HG008972, and UM1 HG006493.

Conflict of Interest: none declared.

References

Boehnke,M. (1986) Estimating the power of a proposed linkage study: a practical computer simulation approach. *Am. J. Hum. Genet.*, **39**, 513–527.

Lander,E. *et al.* (1995) Genetic dissection of complex traits: guidelines for interrupting and reporting linkage results. *Nat. Genet.*, **11**, 241–247.

Ott,J. *et al.* (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.*, **16**, 275–284.

Richards,S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–424.

Sloan-Heggen,C.M. *et al.* (2016) Comprehensive genetic testing in the clinical evaluation of 1119 patients with hearing loss. *Hum. Genet.*, **135**, 441–450.