Bioinformatics, 31(22), 2015, 3682–3684
doi: 10.1093/bioinformatics/btv432
Advance Access Publication Date: 23 July 2015
Applications Note

Genome analysis

GeneVetter: a web tool for quantitative monogenic assessment of rare diseases

Christopher E. Gillies[1], Catherine C. Robertson[1], Matthew G. Sampson[1,*] and Hyun Min Kang[2,*]

[1]Division of Nephrology, Department of Pediatrics and Communicable Diseases, University of Michigan School of Medicine, Ann Arbor, MI, USA and [2]Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 15, 2015; revised on July 3, 2015; accepted on July 19, 2015

Abstract

Summary: When performing DNA sequencing to diagnose affected individuals with monogenic forms of rare diseases, accurate attribution of causality to detected variants is imperative but imperfect. Even if a gene has variants already known to cause a disease, rare disruptive variants predicted to be causal are not always so, mainly due to imperfect ability to predict the pathogenicity of variants. Existing population-scale sequence resources such as 1000 Genomes are useful to quantify the 'background prevalence' of an unaffected individual being falsely predicted to carry causal variants. We developed *GeneVetter* to allow users to quantify the 'background prevalence' of subjects with predicted causal variants within specific genes under user-specified filtering parameters. *GeneVetter* helps quantify uncertainty in monogenic diagnosis and design genetic studies with support for power and sample size calculations for specific genes with specific filtering criteria. *GeneVetter* also allows users to analyze their own sequence data without sending genotype information over the Internet. Overall, *GeneVetter* is an interactive web tool that facilitates quantifying and accounting for the background prevalence of predicted pathogenic variants in a population.
Availability and Implementation: *GeneVetter* is available at http://genevetter.org/
Contact: mgsamps@med.umich.edu or hmkang@umich.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Inaccurate attribution of causality to rare genetic variants found in sequencing studies of subjects with rare diseases can limit the benefits of screening and potentially result in harm (MacArthur *et al.* 2014). For accurate classification of a variant's pathogenicity, investigators typically account for multiple characteristics, such as its allele frequency, conservation score and predicted functional consequences. However, even after careful filtering and application of these parameters to a known set of causal genes, a proportion of non-causal variants are still at risk for being misclassified as causal due to imperfect filtering. For example, it is reported than >1% of general population can be falsely determined to carry causal variants for maturity onset diabetes of the young (MODY), a monogenic form of diabetes (Flannick *et al.*, 2013). Our work is motivated by a study of nephrotic syndrome (NS), a rare disease with more than 20 genes implicated as monogenic causes (McCarthy *et al.*, 2013). Diagnosis of a monogenic cause in an NS patient may be important for personalizing treatment plans. Many studies have quantified the fraction of monogenic cases through targeted sequencing followed by pathogenicity filtering in case-only cohorts (Sadowski *et al.*, 2014). However, applying imperfect variant-filtering procedures will misdiagnose a fraction of truly non-monogenic cases as harboring putative causal variants. In addition, this will result in a non-negligible rate of 'background prevalence' of monogenic NS diagnosis in the unaffected population. Beyond NS, when making genetic diagnosis of any rare diseases in clinical or research settings,

new methods and tools that account for this 'background prevalence' should improve accuracy of this classification.

To facilitate accounting for the background prevalence in sequencing studies of rare diseases, we developed an intuitive web tool, *GeneVetter*. It provides comprehensive analyses and visualizations of putative causal variants when sequencing a set of genes across >2500 individuals in 1000 Genomes (1000G). Users can obtain background prevalence for a specific population and also calculate power and sample size for genetic studies assessing enrichment of case-specific rare and disruptive variants. When sequencing data is already available for a specific cohort, users can apply the same filtering criteria to *GeneVetter* to systematically compare cases with 1000G, without sending any individual-level information over the network.

2 Key features

GeneVetter is a highly interactive web application implemented with *AngularJS*, based on the 1000 Genomes Phase 3 resources annotated by dbNSFP 2.5 and SNPEff 3.5 using GENCODE version 9. The implementation details can be found in the Supplementary Text. In next the section, we describe the key features of *GeneVetter*.

2.1. Quantifying uncertainty in monogenic diagnosis

Given a set of genes known or assumed to causes a monogenic form of disease, *GeneVetter* quantifies the background prevalence that a subject in 1000G carries putative causal variants defined by user-specified criteria. For example, the default variant filtering criteria for genes under a dominant inheritance model considers a variant to be pathogenic if its frequency is less than 0.5% in the Exome Variant Server (EVS) for both European and African Americans and results either in a loss of function, or, a missense change predicted to be damaging by at least two of SIFT, PolyPhen2 and MutationTaster. For genes under a recessive model, the EVS

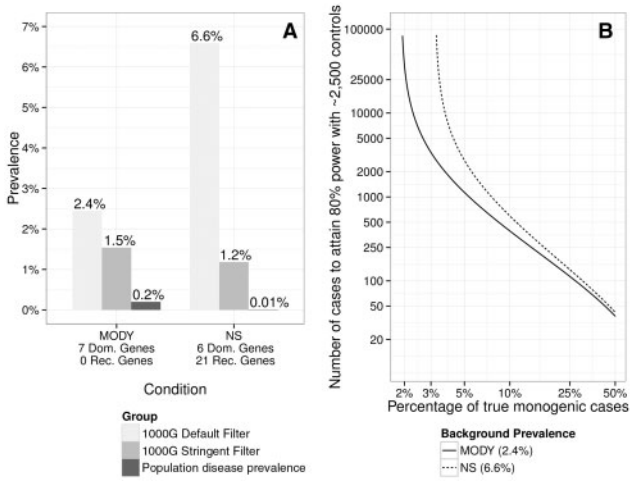

Fig. 1. (**A**) Background prevalence for MODY and NS with monogenic causes estimated by *GeneVetter*. For both diseases, the background prevalence is quantified across ~2500 individuals from 1000G using the default filter and a stringent filter. The stringent filter decreases the AF threshold to 0.1% in the EVS, requires a GERP++ score > 4, removes variants present in > 1% in any individual population of 1000G, and uses the most conserved transcript per gene. (**B**) The number of case samples required to attain 80% power with exome-wide significance ($P = 2.5 \times 10^{-6}$) versus ~2500 controls across different true prevalence of monogenic case (false negative rate = 0.1). Each line represents the required sample size estimated from the background prevalence calculated with default filter in (A) with the same genes

threshold is relaxed to 1%, and two or more variants per gene are required. The filtering criteria are highly customizable with multiple prediction methods available. Through an interactive web interface, *GeneVetter* displays background prevalence not only across all ~2500 subjects in 1000G but also stratified by continental or sub-continental population. It also provides top principal components of genetic ancestry for the carriers of putative causal variants for users to have a detailed understanding of potential effect of population stratification.

Figure 1A illustrates the background prevalence of two rare diseases across known monogenic genes, MODY (7 genes-dominant and 0 genes-recessive mode of inheritance) and NS (6 genes-dominant and 21 genes-recessive mode of inheritance). The striking enrichment of carriers of putative causal variants compared to the known disease prevalence suggests that background prevalence estimates are elevated due to imperfect variant filtering, or incomplete penetrance. When classifying a carrier of putative causal variants as a monogenic case, we expect a similar risk of misclassifying non-monogenic cases as monogenic cases if the populations and sequencing techniques match. Applying a stringent filter is effective to decrease the background prevalence, at the potential expense of decreased sensitivity for identifying truly monogenic cases. The background prevalence quantified by *GeneVetter* can allow users to estimate aggregate relative risk of imperfectly predicted set of causal variants that could help in determining their clinical implications.

2.2. Designing genetic study of rare diseases

GeneVetter enables researchers to devise well-powered targeted sequencing studies. If users have an informed estimate of the proportion of cases carrying putative causal variants across a gene set, *GeneVetter* calculates the number of samples needed to have a specified power to identify significant associations in gene-level burden tests with a specified level of Type I Error. *GeneVetter* uses the background prevalence estimated from 1000G as a default baseline, and calculates the statistical power to identify rare variant associations for a true prevalence of monogenic cases with a specific sample size. The hypothesis test used for the power calculations was a difference in binomial proportions using a z-statistic (see the supplementary Text and *GeneVetter*'s help page). By specifying a small Type I error (e.g. 2.5×10^{-6}), it is also possible to design an exome sequencing study, calibrated by the power to identify a specific set of high priority genes. Figure 1B illustrates an example of the power calculation performed by GeneVetter for MODY and NS.

To further assist with study design, *GeneVetter* also provides a simulation in which a user can specify the background prevalence per gene, number of genes studied and estimated disease prevalence in a case cohort. Users can inspect how these parameters affect the total proportion of subjects correctly and incorrectly classified as monogenic cases, the statistical power or required sample size to identify associations and the false discovery rate.

2.3. Harmonized and secure comparison of sequenced cases with public data

Finally, if users have their own set of sequenced cases, *GeneVetter* allows them to contrast their VCF file to the 1000G populations by applying the same pathogenicity criteria. The filtering is performed in the client's browser using JavaScript. No individual genotype information is sent to the server. Only the variant site list is sent to the server so that *GeneVetter* can obtain the annotations of variants, and the information is immediately discarded after the use. The 'inspect' function summarizes the monogenic prevalence by gene, the

variants that implicate subjects, and all the variants passing the filter criteria. *GeneVetter* also displays a point estimate and 95% confidence interval for the odds ratio of monogenic classification in a user's cases versus the 1000G subjects.

3 Conclusion

Studying monogenic rare diseases while accounting for the background prevalence of putative causal variants provides useful quantitative information for making accurate monogenic diagnosis, calibrating stringency of variant filtering and estimating the effect size from aggregate set of rare variants. *GeneVetter* comprehensively and efficiently achieves this in a highly interactive fashion. While its default settings were chosen carefully, we acknowledge that there is as of yet no optimal filtering strategy. Thus it is also highly customizable, allowing construction of alternative filters. Users can also analyze their sequenced cases with *GeneVetter* and compare with 1000G without transferring individual genotype information over the network. We expect *GeneVetter* to aid clinicians and scientists who may not have time or technical expertise to perform tedious data management tasks, but seek to improve accuracy in implicating variants as causal for their disease of interest.

Acknowledgements

M.S. is supported by 1K08-DK100662 and is a Carl W. Gottschalk Research Scholar of ASN Foundation for Kidney Research.

Conflict of Interest: none declared.

References

Flannick,J. *et al.* (2013) Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.*, **45**, 1380–1385.

MacArthur,D.G. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.

McCarthy,H.J. *et al.* (2013) Simultaneous sequencing of 24 genes associated with steroid-resistant nephrotic syndrome. *Clin. J. Am. Soc. Nephrol. CJASN*, **8**, 637–648.

Sadowski,C.E. *et al.* (2014) A single-gene cause in 29.5% of cases of steroid-resistant nephrotic syndrome. *J. Am. Soc. Nephrol*, **26**, 1279–1289.