RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Population Genetics Identifies Challenges in Analyzing Rare Variants

Henry Richard Johnston,[1] Yijuan Hu,[1] and David J. Cutler[2]*

[1] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, United States of America; [2] Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, United States of America

Received 22 August 2014; Revised 19 November 2014; accepted revised manuscript 20 November 2014.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21881

ABSTRACT: Geneticists have, for years, understood the nature of genome-wide association studies using common genomic variants. Recently, however, focus has shifted to the analysis of rare variants. This presents potential problems for researchers, as rare variants do not always behave in the same way common variants do, sometimes rendering decades of solid intuition moot. In this paper, we present examples of the differences between common and rare variants. We show why one must be significantly more careful about the origin of rare variants, and how failing to do so can lead to highly inflated type I error. We then explain how to best avoid such concerns with careful understanding and study design. Additionally, we demonstrate that a seemingly low error rate in next-generation sequencing can dramatically impact the false-positive rate for rare variants. This is due to the fact that rare variants are, by definition, seen infrequently, making it hard to distinguish between errors and real variants. Compounding this problem is the fact that the proportion of errors is likely to get worse, not better, with increasing sample size. One cannot simply scale their way up in order to solve this problem. Understanding these potential pitfalls is a key step in successfully identifying true associations between rare variants and diseases.
Genet Epidemiol 0:1–4, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: rare variants; whole genome sequencing; error rates

Introduction

The fundamental goal of human genetic research is to identify genetic variants that influence whether an individual develops a particular disease. For the most part, this analysis is done utilizing a statistical framework that asks whether genetic variants appear in diseased individuals more often than expected by chance. The most common manifestation of this process, known as a genome-wide association study, has been successful in identifying thousands of loci which contain common variants (minor allele frequency > 0.05) usually with weak but statistically significant association with disease [Visscher et al., 2012]. However, the loci identified to date may be thought of as "low-hanging fruit," and it is unlikely that a significant number of unidentified common alleles remain for most common diseases. As such, the current paradigm shift involves looking for alleles with relatively large effect sizes but with much lower minor allele frequencies. This is the hunt for rare variants.

In order to study rare alleles, a technology shift must occur. No longer is the genotyping of a million or so variants a useful strategy—instead, whole exomes or whole genomes must be sequenced. In addition, there is virtually no power to detect statistically significant association with an individual rare variant, so testing strategies involving "classes" of variants are being utilized [Wu et al., 2011]. These new approaches bring with them challenges that defy classic genetic intuitions.

Defining Rare Variants

Rare variants can be defined in many ways. Here, we focus on two apparently similar definitions that have profound differences on the behavior of statistical tests. The definitions are so similar, in fact, that one may not realize the important difference between them. First, we begin by focusing on the rarest of rare variants, so-called "singletons," variants seen exactly one time. Even for this case there are two subtly different ways we can define "singleton." For one definition we define a rare variant as a "global singleton" when it occurs in exactly one sample, whether that sample is a case or a control. Thus, it occurs exactly once in a study. For this to be an accurate claim, the researcher must know the positions of all variants in a given study. When this is not the case, a rare variant can instead be a "group singleton" when it is a singleton in cases or in controls or maybe both, without distinction as to whether it appears in both cases and controls. This second definition is often used when the data at hand consists of allelic counts from separate studies, but the underlying allelic identities (i.e., the exact position of the variant in the genome) are unknown to the investigator.

*Correspondence to: David J. Cutler, Department of Human Genetics, Emory University School of Medicine, 343 Whitehead Building, 615 Michael Street, Atlanta, GA 30322, USA. E-mail: dcutler@genetics.emory.edu

Table 1. The expected distribution of global singletons in a case-control study

	Singletons	Nonsingletons
Cases, n_1	$\theta n_1/(n_1+n_2)$	$n_1 - \theta n_1/(n_1+n_2)$
Controls, n_2	$\theta n_2/(n_1+n_2)$	$n_2 - \theta n_2/(n_1+n_2)$

Table 2. The expected distribution of group singletons in a case-control study

	Singletons	Nonsingletons
Cases, n_1	θ	$n_1 - \theta$
Controls, n_2	θ	$n_2 - \theta$

Table 3. Real copy number variant (CNV) data from Mulle et al. 2010. CNVs are an example of a kind of singleton found in the genome. Notice the massive change in *P*-value that occurs when the assumption about whether variants are global or group singletons changes

	Singletons	Nonsingletons
Cases	6	7,539
Controls	1	39,747
P-Value for global singletons	$\sim 9.9 \times 10^{-4}$	
P-Value for group singletons	~ 0.057	

Beginning from a classical result [Watterson, 1975], we let K equal the number of segregating sites and set n to be the total number of alleles studied,

$$E\{K\} = \theta \sum_{i=1}^{n-1} \frac{1}{i} \approx \theta \log(n-1), \tag{1}$$

where θ is the population mutation rate, about 0.001 per nucleotide in humans. This result continues, allowing k_i to be the number of sites with minor allele count i in n sequenced alleles.

$$E\{k_i\} = \theta \left(\frac{1}{i} + \frac{1}{n-i} \right). \tag{2}$$

For singletons, $i = 1$ and this equation simplifies to

$$E\{k_i\} = \theta \left(1 + \frac{1}{n-1} \right) \sim \theta. \tag{3}$$

Notice that the singleton class is nearly independent of sample size (n) for any moderately large study size.

These results allow us to make predictions based on the two slightly different definitions. For global singletons in a case-control study, the expected contingency table looks like Table 1. For group singletons, the expected table looks like Table 2. To understand the import of the difference, we examine the data from a study on CNVs [Mulle et al., 2010]. When the data from this study are entered into our expectation tables, we see a dramatic difference in the P-value of the result (Table 3) due to the change in expectation in Fisher's exact test analysis. The issue of group singletons is a problem that needs to be addressed through investigator understanding. Determining significance in studies as though singletons are true global singletons, when they are, in fact, group singletons, can lead to the publication of false positives.

Unfortunately, the problem gets even slightly worse. The actual expectation for singletons (equation 3) differs slightly by sample size, where n is the number of alleles sampled. Consequently, the expectation for "group singletons" is actually slightly larger for the smaller dataset.

This problem is not specific to the use of singletons. If, instead of singletons, we intend to look at alleles below some frequency threshold, the following logic applies. If we define rare as having a minor allele count less than or equal to r,

$$E\{k_r\} = \theta \sum_{i=1}^{r} \frac{1}{i} + \theta \sum_{i=1}^{r} \frac{1}{n-i}, \tag{4}$$

the first term is dominant and independent of n. The second term is small, but could be substantial for sufficiently small studies, such that more "rare" alleles would be identified in smaller studies. Clearly that is counter to traditional genetic intuition, where the expectation is to identify more variants as sample sizes increase. To demonstrate this, we instead define rare as having a minor allele frequency less than or equal to x, as we are wont to do often in sequencing studies, and we see that

$$E\{k_x\} = \theta \sum_{i=1}^{xn} \frac{1}{i} + \theta \sum_{i=1}^{xn} \frac{1}{n-i}. \tag{5}$$

This scales with sample size (n). Larger sample size here means more rare alleles, as expected intuitively.

This should suggest that defining things as "group singleton" or "group rare" is a bad idea. Unfortunately, it is often done by accident. Combining an available control group with a set of cases analyzed in your lab is the most obvious example. Combining data between studies unless full genotypes for both are available (or an equivalent such as the score statistic at each site plus the covariance matrix) is impossible and should be avoided to ensure accurate results.

Sequencing Error

Next-generation sequencing studies have a significant amount of systematic artifacts that lead to sequencing error. These include mismapping, copy number variants, multiple variants near each other as well as other issues. GATK [DePristo et al., 2011; McKenna et al., 2010] and other filtering techniques are designed to try to eliminate these systemic artifacts. Even after doing so, however, there is surely some residual sequencing error. We start by defining two error rates. Rate e_1 is the probability that a site that is truly a variant is miscalled by the sequencing experiment. This is genotyping error at a polymorphic site. Rate e_2 is the probability that a site that is homozygous for the reference allele is called a variant. This is a false-positive variant call. There errors have different properties and there may well be tradeoffs between them.

Table 4. The fraction of true variants found in introns and intergenic DNA depending on sample size and sequencing error rate

	e_2 error rates			
Sample size	10^{-3}	10^{-4}	10^{-5}	10^{-6}
1	36.6%	85.2%	98.3%	99.8%
10	26.1%	77.9%	97.2%	99.7%
100	5.8%	37.1%	85.5%	98.3%
1,000	1.3%	7.9%	45.1%	89.1%
10,000	1.0%	1.6%	9.9%	51.3%

Table 5. The fraction of true variants found in exons depending on sample size and sequencing error rate

	e_2 error rates			
Sample size	10^{-3}	10^{-4}	10^{-5}	10^{-6}
1	22.4%	74.3%	96.6%	99.7%
10	15.0%	63.8%	94.6%	99.4%
100	3.0%	22.8%	74.6%	96.7%
1,000	0.6%	4.1%	29.1%	80.3%
10,000	0.5%	0.8%	5.2%	34.5%

Error rate e_1 can be estimated by sequencing a single individual and then typing that same individual on a genotyping array. One minus the concordance of those experiments is e_1. This is an easy error rate to estimate, and most papers and algorithms report it as $0.001 < e_1 < 0.01$ [O'Rawe et al., 2013]. This approach slightly overestimates this error rate, since errors in the genotyping array are also included in this estimate.

Estimating e_2 is a somewhat more difficult proposition. To do so, a single individual must be sequenced twice. We define d to be the proportion of sites called differently between the two sequencing experiments.

$$E\{d\} \sim \theta 2e_1 + (1-\theta)2e_2. \quad (6)$$

More directly, let d^* be the proportion of sites that differ, where the site is not known to contain a human variant—essentially any site without a dbSNP entry.

$$E\{d^*\} \sim 2e_2. \quad (7)$$

To understand the true problem, one needs to understand the probability of a true variant existing at a particular site compared to the probability that the site contains a sequencing error that appears to be a variant. If we sequence n individuals, the probability of an error at a given site is approximately

$$\Pr\{Error\} = 1 - (1-e_2)^n \sim ne_2. \quad (8)$$

Sequencing error scales with sample size (n). However, the probability of a true rare variant scales much slower than linear. The probability a site contains a true rare variant with minor allele count less than r is

$$\Pr\{TrueRareVariant\} = \theta \sum_{i=1}^{r-1} \left(\frac{1}{i} + \frac{1}{n-1}\right) \sim \theta \log(r-1) < \theta \log(n). \quad (9)$$

Thus, the probability of a true variant existing at a particular site scales with $\log(n)$. This presents a massive problem as sample sizes get large.

The scale of the problem can be seen in Tables 4 and 5. These tables, derived using equations (8) and (9), show the probability that an identified rare variant is, in fact, real. In introns, in a sample of size 1,000, the e_2 error rates need to be better than (less than) 10^{-6} in order to have 98% of identified variants be real. In exons, which have a lower overall expected mutation rate, that same sequencing error rate leads to only 80% of identified variants being real. A total of 10^{-6} is an excellent error rate, and yet 1/5 of exonic variants will still be false positives. The problem deepens when one understands that per base error rates can differ for a number of reasons such as DNA quality, library construction characteristics, sequencing depths, and others. These things may be correlated with case/control status, which would lead to systemic error. These factors definitely differ between different studies, again suggesting that combining data across studies is a dangerous game.

It should be noted though that all of the models utilized here are based on the assumption that the human population size is constant. The human population is growing [Keinan and Clark, 2012]. As a result, there are more rare sites than expected in the constant neutral model. How many more rare sites are unclear.

Discussion

The study of rare variants is a relatively new pursuit in the realm of human genetics. It is important to note that many of the intuitions that have been finely honed by the study of common variants do not apply, and are in fact incorrect when applied to the study of rare variants. The important things to note on this front are twofold. First, the blind combination of datasets should be avoided. Second, anyone conducting rare variant studies should be aware of their e_2 sequencing error rate. If this rate is not very low, researchers should be skeptical of most rare variant findings. Even with an excellent error rate, investigators should understand that a significant fraction of their newly identified variants is likely to be false positives and temper their enthusiasm appropriately. The study of rare variants is going to be a great challenge for the community, but with proper understanding, measures can be taken to limit errors that might be introduced into published results.

Acknowledgments

We thank Z.S. Qin, R. Yang, Y. Jakubek, A. Locke, K.M. Steinberg, and M. Zwick for informative discussions. This work was supported by a grant from the NIH to D. Cutler (5R01HG003461-05).

References

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M and others. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. London: Nature Publishing Group, Vol. 43, pp. 491–498.

Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336(6082):740–743.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and others. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.

Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC, Sobreira NL, Valle D, Rudd MK, Satten G and others. 2010. Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* 87(2):229–236.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE and others. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 5(3):28.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* 90(1):7–24.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 7(2):256–276.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.