European Journal of Human Genetics (2013) 21, 1027–1030

www.nature.com/ejhg

SHORT REPORT

Empirical power of very rare variants for common traits and disease: results from sanger sequencing 1998 individuals

Martin Ladouceur*,[1,2,6], Hou-Feng Zheng[2,3,6], Celia MT Greenwood[2,4] and J Brent Richards*,[2,3,5]

The optimal study design for identifying rare variants associated with common disease is not yet clear and researchers have to decide whether to prioritize lower sequencing coverage on larger sample sizes, or higher coverage on smaller sample sizes. High-coverage sequencing affords several advantages, such as genotype accuracy and improved identification of very rare variants, but this comes at increased cost. However, the magnitude of the contribution of very rare variants to the statistical power of gene-based association tests is unknown. By using Sanger sequence data on seven genes from 1998 subjects with simulated phenotypes, we provide evidence that excluding very rare variants, in general, reduces the statistical power of rare variant association tests only modestly. However, if the probability of being causal and the effect size of the causal variants are inversely related to the minor allele frequency, then very rare variants do contribute to some power, however the absolute power remains low. As very rare variants constitute the majority of variants identified in sequencing studies, these findings suggest that careful attention need to be placed on the plausible relationship that exist between very rare variants and common disease.
European Journal of Human Genetics (2013) **21,** 1027–1030; doi:10.1038/ejhg.2012.284; published online 16 January 2013

Keywords: rare variants; statistical genetics; statistical power

INTRODUCTION

Many researchers are now focusing efforts on elucidating the role of rare genetic variants in predisposition to common disease.[1] This undertaking usually involves next generation resequencing at high coverage, in order to identify very rare variants with high accuracy.[2] Indeed, the majority of variants identified in large-scale high-depth sequencing studies are rare,[3–6] and as the number of individuals added to a sequencing study increases, there is a disproportionate increase in the number of singletons (variants present in only one individual) identified.[5] The value of these singletons and other very rare variants (defined here as allele frequencies less than 0.125%) in improving power to identify gene-disease associations is currently unclear. In terms of design planning and resource allocation, it is important to know when singletons and other very rare variants have significant impact on power, under which assumptions and statistical models, and when their contribution to power is negligible.

To address this question, we undertook Sanger sequencing of 1998 individuals at seven genes and simulated causal very rare variants to understand their effect on the statistical power of prominent rare variant disease-gene association methods.[7–13]

MATERIAL AND METHODS

Study samples and Sanger sequencing data

The subjects used are a subset of the CoLaus study,[14] and the data consist of Sanger sequences for the exons and flanking regions of seven genes, provided by GlaxoSmithKline (Upper Merion, PA, USA). The sequencing methods have been described previously,[15] and a summary of the genes' characteristics is shown in Table 1.

Simulations

We recently presented a simulation framework using these data to explore parameters potentially influencing rare variant associations with continuous and dichotomous traits.[16] Here, we use the same framework to investigate the impact of the inclusion or exclusion of very rare variants on the power of two recently developed statistical methods for testing rare variants: the variable-threshold approach ,[9] and a variance component regression method, Sequence Kernel Association Test (SKAT).[10] These two methods have good power when compared to other rare variant methods.[16]

Phenotype simulations

Phenotypes were simulated to depend on genetic variants chosen from the complete list of rare variants in a gene, allowing us to assume that very rare variants, such as singletons, can have a deleterious effect on the trait. The phenotypes were simulated to illustrate the power of a variety of commonly held hypotheses about the potential effects of rare variants on traits.

Parameters investigated include: The proportion ($P = 10$, 15, 20, and 30%) of rare variants (MAF $\leq 1\%$) that are causal, combined with effect sizes of the causal rare variants ($\mu = 0.5$, 0.75, 1, 1.25, and 1.5 SD) as the mean effects for a continuous trait; these combinations led to investigation of 20 scenarios. Phenotypes among individuals without any causal genetic variants were assumed to follow a standard normal distribution. A proportion, P, of the rare variants in each gene was randomly chosen to be causal, and phenotype values of individuals carrying at least one rare causal allele was drawn from a normal distribution where the mean is shifted by μ. Two additional scenarios[16]

[1]Research Center of Montreal Heart Institute, Montreal, Quebec, Canada; [2]Department of Epidemiology and Biostatistics, McGill University, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada; [3]Department of Medicine and Human Genetics, McGill University, Montreal, Quebec, Canada; [4]Department of Oncology, McGill University, Montreal, Quebec, Canada; [5]Twin Research and Genetic Epidemiology, King's College London, London, UK
[6]These authors equally contributed to this work.
*Correspondence: Dr M Ladouceur or Dr B Richards, Department of Medicine, McGill University, 3755 Côte Ste-Catherine Road, Montréal H3T 1E2, Québec, Canada H3T 1E2. Tel: +1 514 340 8222; Fax: +1 514 340 7502; E-mail: mladouceur@epimgh.mcgill.ca or brent.richards@mcgill.ca
Received 18 May 2012; revised 12 October 2012; accepted 22 November 2013; published online 16 January 2013

Table 1 Description of the seven genes and count of rare variants per gene

	Total number of variants	*Number of rare variants (MAF < 1%)*	*Median MAF*	*Coding length (base pairs)*	*Number of individuals with different counts of rare variants*				
					0	*1*	*2*	*3*	*≥4*
Gene 1	49	42	2.50E−04	2002	1882	112	4	0	0
Gene 2	103	90	2.54E−04	4094	1707	223	34	4	1
Gene 3	29	27	2.52E−04	1239	1905	74	2	0	0
Gene 4	64	54	5.08E−04	1638	1719	240	9	0	0
Gene 5	68	62	2.54E−04	1963	1771	176	20	2	0
Gene 6	67	54	5.08E−04	2901	1735	223	10	0	1
Gene 7	128	105	5.01E−04	1500	1709	262	24	2	1

explored the assumption that variants with lower MAF have larger effect. In scenario 1, causal variants were sampled based on the inverse of their MAF. The proportion of causal variants was 10%. The effect of each variant was also based on their MAF, with the variant having the lowest MAF receiving an effect of −2.5 SD. The rest of the effect follows equation 1 in Madsen and Browing.[7] Scenario 2 replicates Scenario 1, except that the sampling of the causal variants was uniform. Permutation was used to control for type-1 error in all statistical methods, and power is reported for a type-1 error of 0.05.

RESULTS

We present simulated power results for SKAT and VT. By selectively excluding very rare variants from analysis, eight data sets were created, as detailed in the legends to Figures 1 and 2.

Figure 1a illustrates how the power of SKAT changes across the exclusion criteria. SKAT's power is altered only modestly by exclusion of very rare variants, especially when the effects of causal rare variants are small to moderate. When the variants' effects are larger, the small drop in power owing to removing very rare variants attenuates as the proportion of causal rare variants increases (Figure 1a). Similarly, when assessing the dichotomized phenotype, we did not observe any drop in power when removing very rare variants (Supplementary Figure S1). When we removed all rare variants (as defined by a MAF <0.01), that is, when no variant in the model is associated with the

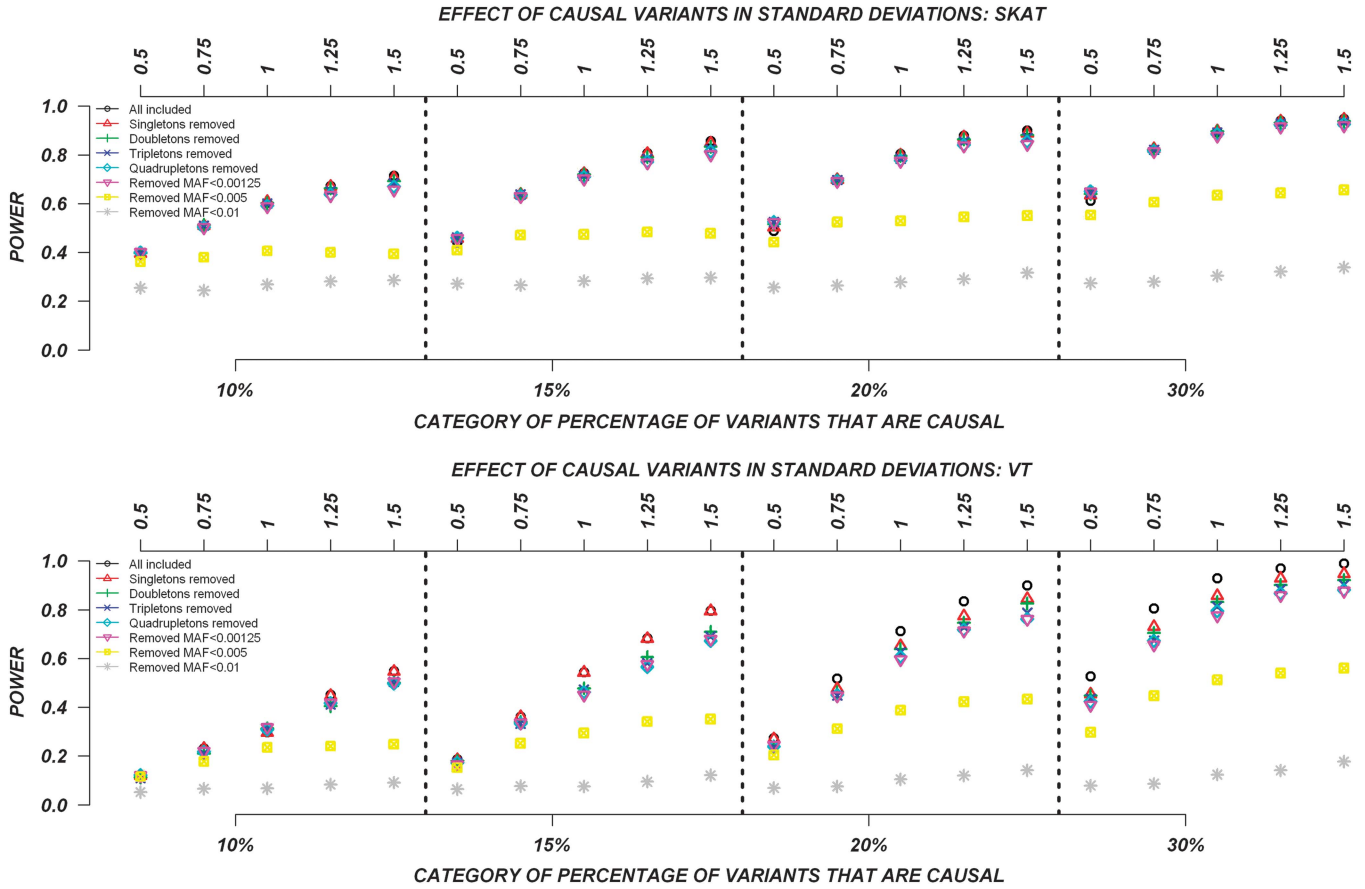

Figure 1 SKAT and VT Continuous Traits: Relationship between effect size, proportion of causal variants and power as rare variants are removed. All causal variants have a deleterious effect. Each box corresponds to a different proportion of causal variants involved in the relationship between rare variants and continuous traits (from left to right, 10, 15, 20 and 30%). On the x-axis, effect sizes are in SD and correspond to the absolute value of the average size effect. By selectively excluding some variants from analysis, eight data sets were created: (1) All variants are included, (2) Singletons are removed, (3) Doubletons and singletons are removed, (4) Tripletons or less are removed, (5) Quadrupletons or less are removed, (6) Variants with MAF <0.00125 are removed, (7) Variants with MAF <0.005 are removed and (8) Variants with MAF<0.01 are removed from the analysis. Top figure illustrates the power under SKAT model, where the bottom figure is draw under VT model.

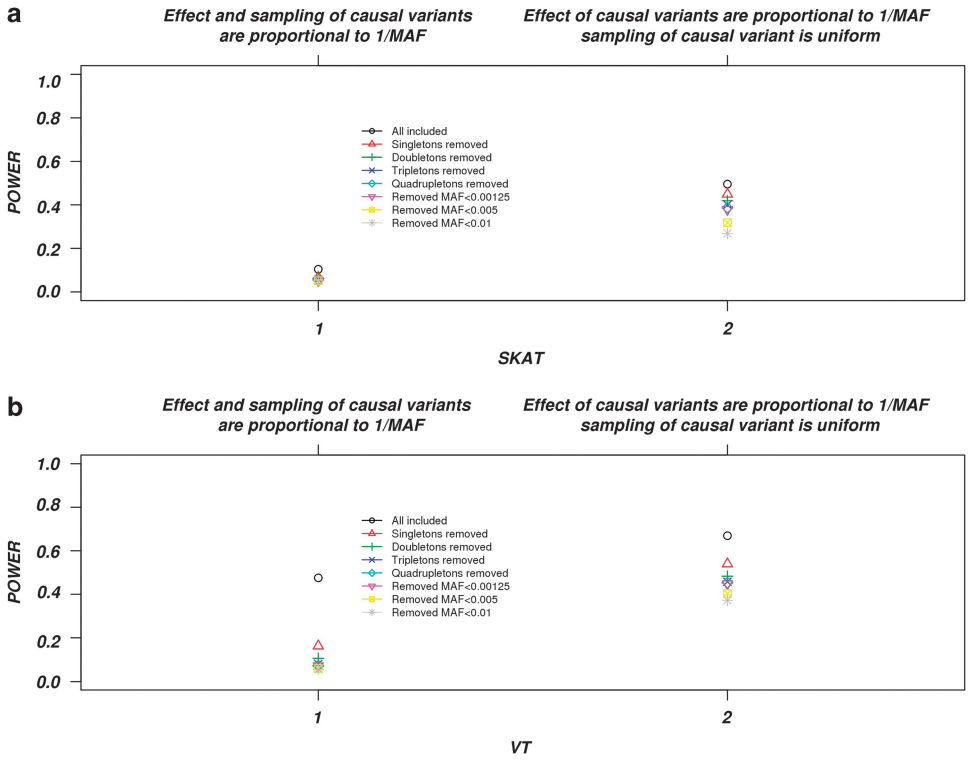

Figure 2 SKAT and VT Continuous Traits: Relationship between the causal variants and their effect size is inversely proportional to the MAF. In Scenario 1, left line, causal variants are sampled based on the inverse of their MAF, and the effect of each causal variant is based on the inverse of their MAF. Scenario 2, right line, is identical to Scenario 1, except that the sampling of the causal variant is uniform, that is, does not depend on MAF. By selectively excluding some variants from analysis, eight data sets were created: (1) All variants are included, (2) Singletons are removed, (3) Doubletons and singletons are removed, (4) Tripletons or less are removed, (5) Quadrupletons or less are removed, (6) Variants with MAF <0.00125 are removed, (7) Variants with MAF <0.005 are removed and (8) Variants with MAF <0.01 are removed from the analysis. Top figure illustrates the power under SKAT model, where the bottom figure is draw under VT model.

trait, the power is very poor in all scenarios. This is expected, as all assigned causal variants were excluded from the analysis.

The power of the collapsing method VT is more affected by the exclusion of rare variants than SKAT. Nevertheless, the loss in power is negligible when singletons are removed, and these are the majority of variants available for analysis. However, power decreased as other thresholds were applied (Figure 1b). For dichotomous phenotypes and VT, power decreases with exclusion of any rare variants (Supplementary Figure S2) by about 20% in some scenarios. The decrease in power for the dichotomous trait analysis is in part explained by the simulation design[16], and because removing all rare variants below a certain cutoff will have an impact on a model that is design to collapse variants below a threshold.

If the rare variants have effect sizes and probability of being causal is inversely proportional to their MAF, power is sensitive to removal of the lowest frequency variants. This difference is more pronounced using VT tests than SKAT (Figure 2). In particular, if both the probability of being causal and the effect size increase for rarer variants, there is a large loss in power associated with exclusion of singletons (Figure 2 first scenario). However, we note that even when including such singletons, absolute power remains low.

DISCUSSION

One of the strengths of our study is the use of Sanger sequencing data, rather than simulated genotyping data on seven genes. And we simulated a variety of phenotype-genotype models on these data. Our choices represent plausible scenarios,[7,9,16] including both constant and allele-frequency dependent effect sizes.

Although the identification of rare variants associated with complex diseases and traits is now underway, our results demonstrate that the inclusion of very rare variants, and particularly singletons, does not always improve power of two popular and powerful rare-variant gene-based association methods. This conclusion does not depend on effect size or the proportion of these variants contributing to the phenotype. However, for dichotomous phenotypes and when rarer variants have stronger effects, power of the VT method may depend substantially on the inclusion of very rare variants of large effect.

In most instances where very rare variants have low effect on power, using lower coverage sequencing on larger sample sizes might be a more suitable allocation of resources. However, this is not always the case. For example, singletons are important when the effect size of causal variants is inversely related to their MAF, but less important if the sampling of causal variants is also inversely related to the MAF. Finally, we note that our study examined seven genes, which are drug targets, and therefore are not necessarily a representative sample of the entire genome.

We recently demonstrated that rare variants have good accuracy even using low coverage sequencing.[17] It is worth noting that Nelson *et al.*[6] did not identify any strong associations with rare variants even among 14 000 people in 202 selected genes. We note however, that a lower depth strategy is not applicable for rare diseases or phenotypically extreme traits, whose etiology may be strongly influenced by very rare or private variants. Our message is not

intended to discourage efforts to identify causal rare variants to exclude them from analysis, but rather, to generate thoughtful discussions about study designs and allocation of resources.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the Canadian Institutes of Health Research (CIHR), the Jewish General Hospital, the Lady Davis Institute, the Fonds de la Recherche en Santé du Québec and the Ministere Développement Économique, Innovation et Exportation du Québec, for funding this research. The authors are indebted to GlaxoSmithKline for the provision of data. The authors thank GlaxoSmithKline and the co-PIs of the CoLaus study, Gerard Waeber and Peter Vollenweider, for the use of the resequencing data, and Drs Matthew R Nelson and Margaret G Ehm for their helpful suggestions as well.

1 Eichler EE, Flint J, Gibson G *et al*: Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; **11**: 446–450.
2 Metzker ML: Sequencing technologies—the next generation. *Nat Rev Genet* 2010; **11**: 31–46.
3 Paola Raska XZ: Rare variant density across the genome and across populations. *BMC Proc* 2011; **5**: S39.
4 Li L, Li Y, Browning SR *et al*: Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* 2011; **6**: e24945.
5 Coventry A, Bull-Otterson LM, Liu X *et al*: Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* 2010; **1**: 131.
6 Nelson MR, Wegmann D, Ehm MG *et al*: An abundance of rare functional variants in 202 drug target genes sequenced in 14 002 people. *Science* 2012; **337**: 100–104.
7 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009; **5**: e1000384.
8 Asimit J, Zeggini E: Rare variant association analysis methods for complex traits. *Annu Rev Genet* 2010; **44**: 293–308.
9 Price AL, Kryukov GV, de Bakker PI *et al*: Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 2010; **86**: 832–838.
10 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011; **89**: 82–93.
11 Liu DJ, Leal SM: Replication strategies for rare variant complex trait association studies via next-generation sequencing. *Am J Hum Genet* 2010; **87**: 790–801.
12 Li Y, Byrnes AE, Li M: To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *Am J Hum Genet* 2010; **87**: 728–735.
13 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008; **83**: 311–321.
14 Firmann M, Mayor V, Vidal PM *et al*: The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc Disord* 2008; **8**: 6.
15 Song K, Nelson MR, Aponte J *et al*: Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 Europeans reveals several rare loss-of-function mutations. *Pharmacogenomics J* 2011; **12**: 425–431.
16 Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB: The empirical power of rare variant association methods: results from sanger sequencing in 1998 individuals. *PLoS Genet* 2012; **8**: e1002496.
17 Zheng H, Ladouceur M, Greenwood C, Richards JB: Effect of genome-wide genotyping and reference panels on rare variants imputation. *J Genet Genomics* 2012; **39**: 545–550.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)