Journal of Human Genetics (2017) 62, 523
© 2017 The Japan Society of Human Genetics All rights reserved 1434-5161/17
www.nature.com/jhg

COMMENTARY

To aggregate or not, that is the question. A commentary on single-nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data

Jurg Ott

Journal of Human Genetics (2017) **62,** 523; doi:10.1038/jhg.2017.7; published online 2 February 2017

In the early days of gene mapping, linkage analysis was carried out with less than two dozen genetic markers.[1] Subsequent development of an abundance of DNA polymorphisms provided a boost to linkage analysis and eventually led to the development of genetic association analysis in the form of case-control studies. These analyses are powerful when variant alleles are relatively frequent, so rare variants are generally disregarded. Newer thoughts have given more credence to the value of rare variants[2] and current efforts, particularly for single-nucleotide variants (SNVs) obtained through sequencing, are directed towards linkage and association analyses including variants with low minor allele frequencies (MAFs).

But how do we carry out association analysis for variants with low MAF? The paper by Yu *et al.*[3] in the current issue of the journal has made a contribution by proposing a particular method for aggregating the effects of multiple SNVs into a single statistic. For candidate disease genes, they propose the proportion of SNVs in a gene (SNVP) as a quantitative marker. On the basis of 46 candidate genes for major depressive disorder and a total of 25 case and control individuals, the authors demonstrate highly significant differences in mean SNVP between cases and controls for some of their candidate genes. They also applied a support vector machine classifier to the SNVP values of all candidate variants and were able to obtain complete separation between case and control individuals. Although aggregating SNVs into a single statistic for a whole gene will not identify a causal variant, such aggregation at least has the potential to pinpoint genes harboring functional SNVs and to rank-order them. The concept of SNVPs makes sense—the higher the proportion of SNVs in a gene, the larger the chance that it is disrupted. However, the report discussed here[3] only addresses the concept of SNVPs and does not contain proof that genes with high SNVPs are in fact disease genes.

Aggregating SNVs over a gene has previously been proposed. In the simplest approach, genotypes are collapsed over a gene[4] and an individual is assigned a code of 1 if an SNV is present in the gene and a code of 0 otherwise. One may then test whether the proportion of individuals with a code of 1 is significantly different between cases and controls.[4] These types of aggregating variants over genes are now known as *burden tests*,[5] and various forms of these tests have been developed.[6]

Aggregating information over multiple neighboring variants had been developed before collapsing and burden tests were introduced: SNV-specific test statistics or other characteristics may be summed over a sequence of SNVs and lead to the concept of *scan statistics*.[7] The SNVP is also a sum over variants, as are statistics in burden tests and analogous approaches. The authors, Yu *et al.*,[3] have now added a new wrinkle to this growing family of tests.

A somewhat different form of aggregation over SNVs is based on the Hamming distance (HD). In a comparison of a sequence of SNVs between two individuals, the relative HD as the proportion of SNVs that are different in the two individuals has been shown to provide top ranks to known disease variants.[8,9]

CONFLICT OF INTEREST

The author declares no conflict of interest.

J Ott is at Laboratory of Statistical Genetics, Rockefeller University, New York, NY, USA
E-mail: ott@rockefeller.edu

1 Ott, J., Schrott, H. G., Goldstein, J. L., Hazzard, W. R., Allen, F. H. Jr, Falk, C. T. *et al.* Linkage studies in a large kindred with familial hypercholesterolemia. *Am. J. Hum. Genet.* **26**, 598–603 (1974).

2 McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).

3 Yu, C., Baune, B. T., Licinio, J. & Wong, M.-L. Single-nucleotide variant proportion in genes: a new concept to explore major depression based on DNA sequencing data. *J. Hum. Genet.* **62**, 577–580 (2017).

4 Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).

5 Li, B., Liu, D. J. & Leal, S. M. Identifying rare variants associated with complex traits via sequencing. *Curr. Protoc. Hum. Genet.* Ch.**1**, 1–26 (2013).

6 Santorico, S. A. & Hendricks, A. E. Progress in methods for rare variant association. *BMC Genet.* **17**(Suppl 2), 6 (2016).

7 Hoh, J. & Ott, J. Scan statistics to scan markers for susceptibility genes. *Proc. Natl Acad. Sci. USA* **97**, 9615–9617 (2000).

8 Imai, A., Nakaya, A., Fahiminiya, S., Tetreault, M., Majewski, J., Sakata, Y. *et al.* Beyond homozygosity mapping: family-control analysis based on Hamming distance for prioritizing variants in exome sequencing. *Sci. Rep.* **5**, 12028 (2015).

9 Imai, A., Kohda, M., Nakaya, A., Sakata, Y., Murayama, K., Ohtake, A. *et al.* HDR: a statistical two-step approach successfully identifies disease genes in autosomal recessive families. *J. Hum. Genet.* **61**, 959–963 (2016).