frontiers in Genetics

EDITORIAL
published: 08 April 2015
doi: 10.3389/fgene.2015.00133

Associating rare genetic variants with human diseases

Qunyuan Zhang *

Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO, USA

Keywords: rare genetic variants, human diseases, association, statistical analysis methods, burden tests, non-burden tests

Genome researches have revealed that a large portion (over 50%) of genetic variants on human chromosomes are rare variants (RVs) with extremely low allele frequency (usually defined as less than 1%) in populations. In recent years, advances of DNA genotyping and sequencing technologies have been facilitating the discovery of RVs, and the association between RVs and human diseases is of rapidly growing interest in understanding genetic and molecular mechanism of both common and rare diseases (Cirulli and Goldstein, 2010; Gibson, 2012).

As a potential source of contribution to the "missing heritability" that cannot be explained by common variants in most SNP-array based genome wide association studies (GWAS), RVs have been identified to be associated with some human diseases or disease-related complex traits, such as cholesterol level (Cohen et al., 2004), hypertriglyceridemia (Johansen et al., 2010), autoimmune disease (Hunt et al., 2012), and Alzheimer's disease (Lord et al., 2014). In spite of these findings, it is still quite challenging to identify the RV association for many diseases and traits, due to the rarity of RVs in populations.

When allele frequencies become very low, the odds of observing RVs in study samples will be small. As a result, the lack of variation in the observed data usually makes the statistical test of association significantly underpowered, which has been widely recognized as the major issue in most RV association studies. To improve the power, numerous statistical methods have been developed, investigated, and used in recent years (Bansal et al., 2010; Basu and Pan, 2011; Ladouceur et al., 2012; Dering et al., 2014; Lee et al., 2014), mostly based on a hypothesis that multiple RVs within a genetic unit (usually a gene) may function in a similar way and collectively contribute to a disease (such phenomena have been observed in many diseases). These methods can be categorized into burden test, non-burden test, and unified test. When burden tests, e.g., CMC (Li and Leal, 2008) and WSS (Madsen and Browning, 2009), are designed for detecting the association of a genotypic "burden" score summarized from a set of RVs, non-burden tests keep individual RVs as individual variables and evaluate whether at least one of multiple RVs is associated with the trait (Wu et al., 2011; Kim et al., 2014; Wang, 2014), and unified tests provide a hybrid analysis combining both burden and non-burden tests (Lee et al., 2012; Sun et al., 2013). In general, burden tests are more powerful when the portion of causal variants increases, non-burden tests may outperform when only a small portion of variants are causal, and unified tests provide an optimized balance between burden and non-burden methods (Lee et al., 2012). One key feature of these methods is that they test the collective effect (not individual effects) of multiple RVs as an entire group, therefore, once the association of a group of RVs is identified, further analyses are still required to determine which one or ones in the group cause the association. Another limitation is that although these methods provide useful tools for association analysis, they usually lack the estimation of heritability of RVs; further analyses of heritability using appropriate methods, e.g., bootstrap-sample-split algorithms (Liu and Leal, 2012), are still warranted.

The main benefit from collective tests is that statistical power can be substantially increased if the majority (or a significant portion) of a set of RVs have effects on a trait. However, the power can be compromised by neutral or "noise" variants (i.e., the variants with no effects) mingled in

OPEN ACCESS

Edited by:
Rongling Wu,
Pennsylvania State University, USA

Reviewed by:
Li Zhang,
University of California, San Francisco, USA
Kwangmi Ahn,
National Institute of Mental Health, USA

***Correspondence:**
Qunyuan Zhang,
qunyuan@wustl.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *15 January 2015*
Accepted: *19 March 2015*
Published: *08 April 2015*

Citation:
Zhang Q (2015) Associating rare genetic variants with human diseases. Front. Genet. 6:133. doi: 10.3389/fgene.2015.00133

analysis. To deal with this issue, many bioinformatics databases and tools (Adzhubei et al., 2010; Wang et al., 2010; Preeprem and Gibson, 2014) have been used to annotate or predict the functions of RVs, and the variants that are unlikely to have functions (e.g., synonymous variants) are usually excluded. For those RVs kept in an analysis after the exclusion based on bioinformatics prediction, since their contributions to a trait could be very different, many methods have been proposed to assign different weights to individual RVs when building a collective test. Weights for individual RVs can be determined by, for example, allele frequency (Madsen and Browning, 2009), genotyping score, annotation score, and many statistics or probabilities estimated from the experimental and/or public data. Particularly, adaptive weighting methods (Lin and Tang, 2011; Pan and Shen, 2011; Zhang et al., 2011) even allow weights to be learned from an initial association analysis using the observed genotype and phenotype data; usually, these methods can improve power at a cost of computational burden, especially when sample sizes are large.

In a collective test, because RVs are tested set by set (not variant by variant), how to select a RV set is critical and will significantly affect the power. Although typically RVs are grouped and selected by gene, there are many other ways to determine the RV sets, for example, by chromosomal region, by exon, by functional pathway, etc. The allele frequency cutoff can be fixed (e.g., <0.01) or a variable threshold method (Price et al., 2010) can be adopted. Since there is no a gold standard for RV grouping and selection, in practice it may need to try multiple ways based on different hypotheses to optimize an analysis.

One of the most popular experiment designs for the RV association analysis is to compare two groups of subjects, such as case and control groups, or two groups differentiating in a quantitatively measurable trait (e.g., low vs. high blood pressure). Analysis of data from such designs can be performed in either a simple way that tests the enrichment of RVs in one group (vs. another group), or a more sophisticate way that models the group assignment as a binary outcome variable dependent on individual and/or summarized RV variables as well as other confounding covariates (e.g., age and gender) if necessary. Although most designs and statistical methods were originally introduced for binary outcomes from unrelated subjects, many of them have been extended to quantitative traits and family data (Fang et al., 2012; Chen et al., 2013; De et al., 2013; Ionita-Laza et al., 2013; Zhang et al., 2014).

Since the potential of power improvement via statistical methods is limited, the ultimate and most effective way of increasing power yet relies on larger sample sizes, which can be realized by either increasing the sample size in a single experiment, or combining the results from multiple experiments into a meta-analysis using recently developed methods (Hu et al., 2013; Lee et al., 2013; Feng et al., 2014; Liu et al., 2014). Once an adequate power is achieved and a significant, solid statistical association between RVs and a disease has been identified, the associated RVs can be very useful in many ways, such as biomarker design, biological function validation, molecular mechanism discovery, drug development, etc.

Acknowledgments

I would like to thank all the authors, reviewers, editors, managers, and publishers who have supported and contributed to this special issue on the research topic Identification of Rare Genetic Variants Contributing to Human Diseases.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* 11, 773–785. doi: 10.1038/nrg2867

Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* 35, 606–619. doi: 10.1002/gepi.20609

Chen, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* 37, 196–204. doi: 10.1002/gepi.21703

Cirulli, E. T., and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11, 415–425. doi: 10.1038/nrg2779

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., and Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305, 869–872. doi: 10.1126/science.1099870

De, G., Yip, W. K., Ionita-Laza, I., and Laird, N. (2013). Rare variant analysis for family-based design. *PLoS ONE* 8:e48495. doi: 10.1371/journal.pone.0048495

Dering, C., Konig, I. R., Ramsey, L. B., Relling, M. V., Yang, W., and Ziegler, A. (2014). A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required. *Front. Genet.* 5:323. doi: 10.3389/fgene.2014.00323

Fang, S., Sha, Q., and Zhang, S. (2012). Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet. Epidemiol.* 36, 499–507. doi: 10.1002/gepi.21646

Feng, S., Liu, D., Zhan, X., Wing, M. K., and Abecasis, G. R. (2014). RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* 30, 2828–2829. doi: 10.1093/bioinformatics/btu367

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Hu, Y. J., Berndt, S. I., Gustafsson, S., Ganna, A., Genetic Investigation of Anthropometric Traits Consortium, Hirschhorn, J., et al. (2013). Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *Am. J. Hum. Genet.* 93, 236–248. doi: 10.1016/j.ajhg.2013.06.011

Hunt, K. A., Smyth, D. J., Balschun, T., Ban, M., Mistry, V., Ahmad, T., et al. (2012). Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. *Nat. Genet.* 44, 3–5. doi: 10.1038/ng.1037

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet.* 21, 1158–1162. doi: 10.1038/ejhg.2012.308

Johansen, C. T., Wang, J., Lanktree, M. B., Cao, H., McIntyre, A. D., Ban, M. R., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* 42, 684–687. doi: 10.1038/ng.628

Kim, S., Pan, W., and Shen, X. (2014). Penalized regression approaches to testing for quantitative trait-rare variant association. *Front. Genet.* 5:121. doi: 10.3389/fgene.2014.00121

Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M., and Richards, J. B. (2012). The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet.* 8:e1002496. doi: 10.1371/journal.pgen.1002496

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet*. 95, 5–23. doi: 10.1016/j.ajhg.2014.06.009

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet*. 91, 224–237. doi: 10.1016/j.ajhg.2012.06.007

Lee, S., Teslovich, T. M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet*. 93, 42–53. doi: 10.1016/j.ajhg.2013.05.010

Li, B., and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet*. 83, 311–321. doi: 10.1016/j.ajhg.2008.06.024

Lin, D. Y., and Tang, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet*. 89, 354–367. doi: 10.1016/j.ajhg.2011.07.015

Liu, D. J., and Leal, S. M. (2012). Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am. J. Hum. Genet*. 91, 585–596. doi: 10.1016/j.ajhg.2012.08.008

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nat. Genet*. 46, 200–204. doi: 10.1038/ng.2852

Lord, J., Lu, A. J., and Cruchaga, C. (2014). Identification of rare variants in Alzheimer's disease. *Front. Genet*. 5:369. doi: 10.3389/fgene.2014.00369

Madsen, B. E., and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 5:e1000384. doi: 10.1371/journal.pgen.1000384

Pan, W., and Shen, X. (2011). Adaptive tests for association analysis of rare variants. *Genet. Epidemiol*. 35, 381–388. doi: 10.1002/gepi.20586

Preeprem, T., and Gibson, G. (2014). SDS, a structural disruption score for assessment of missense variant deleteriousness. *Front. Genet*. 5:82. doi: 10.3389/fgene.2014.00082

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet*. 86, 832–838. doi: 10.1016/j.ajhg.2010.04.005

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol*. 37, 334–344. doi: 10.1002/gepi.21717

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 38:e164. doi: 10.1093/nar/gkq603

Wang, X. (2014). Firth logistic regression for rare variant association tests. *Front. Genet*. 5:187. doi: 10.3389/fgene.2014.00187

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet*. 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Zhang, Q., Irvin, M. R., Arnett, D. K., Province, M. A., and Borecki, I. (2011). A data-driven method for identifying rare variants with heterogeneous trait effects. *Genet. Epidemiol*. 35, 679–685. doi: 10.1002/gepi.20618

Zhang, Q., Wang, L., Koboldt, D., Boreki, I. B., and Province, M. A. (2014). Adjusting family relatedness in data-driven burden test of rare variants. *Genet. Epidemiol*. 38, 722–727. doi: 10.1002/gepi.21848

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.