Editorial

Rare variants: data types and analysis strategies

Chayanika Goswami[1,2], Amrita Chattopadhyay[2], Eric Y. Chuang[2,3,4]

[1]Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei; [2]Centre of Genomic and Precision Medicine, National Taiwan University, Taipei; [3]Department of Electrical Engineering, National Taiwan University, Taipei; [4]China Medical University, Taichung

Correspondence to: Eric Y. Chuang, ScD, EMBA. Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei. Email: chuangey@ntu.edu.tw.

Submitted Apr 06, 2021. Accepted for publication Apr 25, 2021.
doi: 10.21037/atm-21-1635
View this article at: http://dx.doi.org/10.21037/atm-21-1635

Rare variants are defined as single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) of less than 0.01. They often have larger phenotypic effects in comparison to low-frequency (less common) (0.01< MAF <0.05) or common (MAF >0.05) disease-associated SNPs (1). Genome wide association studies (GWASs) have been extensively used to investigate the underlying genetic etiology of complex diseases and quantitative traits. The GWAS catalog (https://www.ebi.ac.uk/gwas/) is a repository of more than 70,000 identified disease-associated variants. It provides numerous novel clues to disease biology, thus improving our knowledge from a few positionally cloned loci to several thousands of replicated associations.

Despite such findings, the genetic predisposition toward many complex disease traits is left unexplained, even for diseases where large GWAS meta-analyses have been performed. Overall, <24% of the heritability of complex diseases has been accounted for (2). Furthermore, translating such information into functional understanding or therapeutic treatment is still a decades-long journey. Among several hypotheses regarding why there is still missing heritability, one is that GWASs primarily identify common variants (3). In contrast, studies of low-frequency or rare variants can provide an enhanced number of insights into disease risk and trait variability.

The success of rare variant studies highly depends on the design scheme of the experiment. The choice of data and its preprocessing determines a viable start for the discovery of crucial rare variants. Low-depth whole genome sequencing (WGS) is a preferred option in rare variant studies, as deep WGS is often expensive (4). Moreover, variant detection and disease detection are quite achievable with low-depth sequencing, if the sample is large. Exome sequencing is another option; while it only targets the coding region of the genome, as studies show, many Mendelian disorders have been identified through it (3). Despite the fact that many GWAS loci lie in the non-coding region, concentrating a study on a high-value region of the genome still proves to be worthwhile for rare variant studies, keeping in mind the cost of WGS. Targeted-region sequencing has also proven to be effective (5). The discovery of common variants associated with known complex diseases in GWASs has paved the way for targeted-region sequencing and discovery of rare variants (5). It is also a cost-effective approach and is promising for the identification of rare variants.

Customized genotyping arrays are a cost-effective alternative to sequencing. Such chips include both common and rare variants, with common variants replicating the original GWAS signals, thereby enabling fine mapping of rare variants. Extreme-phenotype sampling is a strategy of smart sampling for rare variant studies. Preferential selection of the most likely informative samples while designing arrays can greatly reduce the sequencing budget and increase the association power. Sampling those individuals with known disease phenotypes and risk factors enriches the arrays with rare variants (2).

As rare variants are numerous and are less closely correlated with each other in comparison to common variants, they suffer from multiple testing burden. Rare variant association tests further suffer from a decrease in statistical power due to the rarity of individuals carrying these variant alleles. Hence, rare variant association is confirmed by combining multiple variants within genetic units of association, which are defined by gene annotations,

genomic coordinates, or functional characterization. Burden tests, adaptive burden tests, variance component tests, and combined tests are some of the gene-based tests used frequently for rare variant association studies (2).

Burden tests create genetic scores by collapsing the rare variant count. The key principle behind them is to aggregate the information contained in multiple genotypes of one sample into a burden score that can be easily used for association, and they assume that all rare variants that are causal and trait-associated have the same intensity of effect. The cohort allelic sums test (CAST) is one such burden test, and is available as an R package (6). Similar to the single-SNP analysis, χ^2 and Fisher exact tests can also be used for burden testing (7), depending on the dataset tested. However, a limitation of burden tests is that they assume that all variants influence the phenotype in the same direction.

Adaptive burden tests are more robust than burden tests, being applied with fixed thresholds. They remove the limitations of a burden test and allow the presence of null, trait-decreasing, and trait-increasing variants. Adaptive burden tests are computationally intensive, as they require permutation for the computation of P values. They also make use of regression coefficient for each variant, to be used as weights.

Variance component tests allow a mixture of effects across rare variant sets, including both protective and risk variants, with varying magnitudes of effect sizes. The sequence kernel association test (SKAT) employs this method (8). It can be applied for both quantitative and binary traits. SKAT-O is based on a blend of the burden test and the variance component test, commonly known as a combination test (9). It allows for a more flexible framework in terms of score statistics, leading to an optimal combination of efficient computations. cSKAT (10) can be used to optimize SKAT statistic over multiple potentially relevant SNV annotations. It is powerful for larger sample size ($N \geq 5,000$) and correctly specified SNV weights.

A recent study proposes the Bayes Factor method for rare variant association test in sequencing data. It has informative priors which shows sensitivity to rare variant count differences in binary studies or allelic distribution differences or both. Although it could be applied to unbalanced case-control study designs but it hasn't been tested for different population structures (11). Adaptive hierarchically structured variable selection (HSVS-A) (12) is another newly proposed method which is powerful than both burden test and variance-component tests for continuous phenotypes. It can be applied to both set-based and region-based analysis. It automatically controls the type I error rates and can produce individual effect estimates for rare variants. Association test for rare variants based on algebraic statistics (ASRV) is a novel method to test association when the causal variants has effects in different directions (13). Single variant association tests such as Transmission Disequilibrium Tests (TDTs) (14) or Family-based Association Tests (FBATs) (15) that are robust against genetic confounding can be applied in family-based association studies. Aggregated Cauchy Association Test (ACAT) (16) is a set-based association test for sequencing studies. It is computationally efficient and requires only P values for association test between a trait and a variant-group. RVFam is an R package providing tools for testing association between rare variants and continuous traits, binary traits or survival measured in family samples (17), but outperformed by generalized linear model (GLM) (18) and Firth test (19) which do not account for sample-relatedness. A hybrid strategy using GLM for gene-based tests and Firth test otherwise with family data for rare variant association analyses of binary outcome proves to be computationally efficient without variant filtration. The Bayesian rare variant Association Test using Integrated Nested Laplace Approximation (BATI) test integrated in the rare-variant Genome Wide Association Study (rvGWAS) framework (20) includes both categorical and numerical variant characteristics as covariates for rare variant association test and shows powerful analysis in case of loss-of-function variants.

A pathway-based approach or multi-set testing for rare variant association test shows increase in statistical power when the subsets of genes such as exons, introns or gene windows contains fewer variants overall and may also improve potential disease-etiology elucidation (21). Copula-based Joint Analysis of Multiple Phenotypes (C-JAMP) is a single-marker association test, implemented as an R package, which uses a joint model of various phenotypes and variants or other covariates and is powerful for causal variants with large effect sizes (22). Quantitative Phenotype Scan Statistic (QPSS) has an advantage of localizing genomic regions of rare quantitative-phenotype-associated variant group by refining a known region of interest using variant annotation (23).

For rare variant association tests in non-coding regions such as introns, promoters, enhancers or silencers, a sliding window approach can be used to scan the genome, especially in WGS studies, as there have been very few studies in such

regions. SAIGE-GENE is a scalable generalized mixed model region-based association test, used in exome-wide and genome-wide region-based analysis for large sample data (N>40,000). It can also work with unbalanced case-control ratios for binary traits and control the type I error rates well (24). For region-based rare-variant association studies in WGS data, GECS helps in estimating the significance thresholds, with FWER controlled at 5%. For single-marker analysis studies, a significance threshold of $\alpha = 5.0\times10^{-8}$ had been set based on previous studies. But GECS shows the threshold to be much stringent with $\alpha = 2.95\times10^{-8}$ (25).

Other issues for rare variant studies include population stratification and genotype imputation. The former often acts as a confounder in such studies and should be adjusted before proceeding. Principal component analysis and linear-mixed models are commonly used for this purpose; however, it is not clear how effective they are for rare variants. Genotype imputation, on the other hand, negatively affects rare variant studies, as the imputation accuracy decreases with lower MAFs, leading to removal of rare variants in the quality-check step (2). The introduction of a hybrid reference panel may help resolve this issue, leading to rare variants being imputed with higher accuracy (1).

For a true positive association of rare variants with a disease, it is important to replicate the association in a large number of samples, often relying on sequencing or genotyping. Follow-up studies targeting high-priority variants in multiple samples can be beneficial once the initial study has proven informative (2). Further experiments linking the rare variants to molecular and cellular functions can be carried out once clear evidence of an association with disease is established.

Acknowledgments

We thank Melissa Stauffer, PhD. for English editing our manuscript.
Funding: Center for Biotechnology, National Taiwan University: GTZ300.

Footnote

Provenance and Peer Review: This article was a standard submission to the journal. The article did not undergo external peer review.

Conflict of Interests: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-21-1635). All authors report funding from National Taiwan University: GTZ300.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Chattopadhyay A, Lu TP. Overcoming the challenges of imputation of rare variants in a Taiwanese cohort. Transl Cancer Res 2020;9:4065-9.
2. Lee S, Abecasis GR, Boehnke M, et al. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 2014;95:5-23.
3. Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011;12:745-55.
4. Zhou Q, Moser T, Perakis S, et al. Untargeted profiling of cell-free circulating DNA. Transl Cancer Res 2018;7:S140-52.
5. Morrison AC, Bis JC, Hwang SJ, et al. Sequence analysis of six blood pressure candidate regions in 4,178 individuals: the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) targeted sequencing study. PLoS One 2014;9:e109155.
6. Nicolae DL. Association Tests for Rare Variants. Annu Rev Genomics Hum Genet 2016;17:117-30.
7. Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 2010;18:111-7.
8. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 2011;89:82-93.
9. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics

2012;13:762-75.
10. Posner DC, Lin H, Meigs JB, et al. Convex combination sequence kernel association test for rare-variant studies. Genet Epidemiol 2020;44:352-67.
11. Xu J, Xu W, Briollais L. A Bayes factor approach with informative prior for rare genetic variant analysis from next generation sequencing data. Biometrics 2021;77:316-28.
12. Yang Y, Basu S, Zhang L. A Bayesian hierarchically structured prior for rare-variant association testing. Genet Epidemiol 2021;45:413-24.
13. Meng J, Zhu W, Li C, et al. A novel association test for rare variants based on algebraic statistics. J Theor Biol 2020;493:110228.
14. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16.
15. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 2006;7:385-394.
16. Liu Y, Chen S, Li Z, et al. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet 2019;104:410-21.
17. Chen MH, Yang Q. RVFam: An R package for rare variant association analysis with family data. Bioinformatics 2016;32:624-6.
18. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med 2002;21:2409-19.
19. Firth D. Bias Reduction of Maximum Likelihood Estimates. Biometrika 1993;80:27.
20. Susak H, Serra-Saurina L, Demidov G, et al. Efficient and flexible Integration of variant characteristics in rare variant association studies using integrated nested Laplace approximation. PLoS Comput Biol 2021;17:e1007784.
21. Fore R, Boehme J, Li K, et al. Multi-Set Testing Strategies Show Good Behavior When Applied to Very Large Sets of Rare Variants. Front Genet 2020;11:591606.
22. Konigorski S, Yilmaz YE, Janke J, et al. Powerful rare variant association testing in a copula-based joint analysis of multiple phenotypes. Genet Epidemiol 2020;44:26-40.
23. Katsumata Y, Fardo DW. Quantitative phenotype scan statistic (QPSS) reveals rare variant associations with Alzheimer's disease endophenotypes. BMC Med Genet 2020;21:106.
24. Zhou W, Zhao Z, Nielsen JB, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat Genet 2020;52:634-9.
25. Kanoungi G, Nothnagel M, Becker T, et al. The exhaustive genomic scan approach, with an application to rare-variant association analysis. Eur J Hum Genet 2020;28:1283-91.

Cite this article as: Goswami C, Chattopadhyay A, Chuang EY. Rare variants: data types and analysis strategies. Ann Transl Med 2021;9(12):961. doi: 10.21037/atm-21-1635