Pathogenicity of De Novo Rare Variants
Challenges and Opportunities

Arya Mani, MD

Human molecular genetics has played a critical role in the discovery of novel disease pathways and identification of new targets for therapeutic development. The most significant advantage of this scientific field is its unique potentials to establish causal links between germline mutations and human diseases. This in turn has led to the identification of most relevant targets in humans for development of potent therapeutics. This general concept pertains mainly to single gene or so-called Mendelian disorders, which are largely caused by mutations that alter a protein structure or function and have sufficient power to independently cause disease. Before the advent of high-throughput sequencing, these variants were largely identified by positional cloning. Regardless of the tools used for their discovery, disease causality of Mendelian variants is primarily established by close to perfect segregation of the disease alleles with the trait in family-based studies. A major benefit of family-based studies is the common genetic background of the studied subjects, which allows circumventing the problem of population stratification. Selective pressures in direct relationship to the effect size and severity of disease alleles determine the allele frequencies. For instance, fitness-related traits are highly subjected to natural selection and are caused by variants with much lower allele frequencies compared with those that underlie late-onset diseases.[1] In general, disease allele frequencies of Mendelian traits are low and at a fraction of their prevalence. With the advent of high-throughput sequencing, the ability to identify rare Mendelian variants has dramatically increased. The reducing cost of sequencing and its increased throughput have turned whole-exome sequencing and whole-genome sequencing to increasingly attractive genetic tools for Mendelian traits. The modern tools of whole-exome sequencing or whole-genome sequencing have facilitated discovery of novel rare variants for Mendelian disorders with previously unknown genetic causes. These, in turn, have led to the discovery of novel disease pathways that may facilitate drug development in the near future.

See Article by Paludan-Müller et al

In the opposite spectrum of rare Mendelian variants are common variants with minor allele frequencies >5% and effect sizes that are small and insufficient to independently cause disease. Common variants reside mainly in noncoding regions of the genome, are defined by their disease association (not necessarily causation), and are underpowered to show disease causality. Since the completion of Human Genome Project, there has been an exponential increase in the number of common variants predisposing to complex traits, largely identified by genome-wide association studies (GWAS). One limitation of GWAS is the population stratification, evidenced by the lack of reproducibility of many discovered loci in independent populations. Most importantly, however, the identified GWAS variants are not the actual functional variants but linked to them through linkage disequilibrium. Further, the small effect size of the variants confounds the interpretation of their functional consequences. The identified common genetic variants account only for a fraction of disease heritability, a phenomenon broadly branded as missing heritability. This problem is largely because of the limitations of GWAS in detecting low-frequency disease alleles, even when large sample sizes are used.

The missing heritability in GWAS is believed to be accounted for by low allele frequency variants or variants with small effect size. This concept has generated increased interest for the interrogation of low-frequency (minor allele frequencies 1% to 5%) and rare variants (minor allele frequencies <1%). One approach for discovery of these variants is the large-scale statistical imputation from dense reference panels, which enable inference for unobserved genotypes. This approach, however, is underpowered to handle rare variants. Consequently, specialized chips have been developed to assess large number of rare and low-frequency variants. These include Immunochip, which has rather an incomplete coverage of the low-frequency and rare variants[2] and later developed custom arrays that contain greater number of rare coding variants such as the Metabochip and UK Biobank Axiom and Illumina HumanExome BeadChip Arrays. Regardless of their sizes, these arrays test only definite variants and, hence, have shown modest success in identifying novel rare variants for diseases.

The reducing cost and the enormous power of whole-exome sequencing and whole-genome sequencing in identifying novel variants has made these platforms as most attractive. These technologies were first used to identify unknown disease genes for Mendelian traits. Their widespread use soon unraveled the unpredicted abundance of novel rare variants in healthy individuals.[3] The exome sequencing in thousands of people showed that each individual in average has >20 heterozygotes and 1 homozygote novel loss-of-function variant and that almost every single gene from existing disease genes to those that encode drug targets harbor rare heterozygous loss-of-function variants.[4] The excess of rare variants identified by whole-exome sequencing is explained by the explosive human

From the Department of Internal Medicine, Department of Genetics, Yale School of Medicine, New Haven, CT.

Correspondence to Arya Mani, MD, Yale Cardiovascular Research Center, 300 George St, New Haven, CT 06511. E-mail arya.mani@yale.edu

(*Circ Cardiovasc Genet*. 2017;10:e002013.
DOI: 10.1161/CIRCGENETICS.117.002013.)

***Circ Cardiovasc Genet* is available at**
http://circgenetics.ahajournals.org
DOI: 10.1161/CIRCGENETICS.117.002013

population growth.[5] The effect sizes of these variants are significantly larger than those of common variants but not to the extent to be independently causal. Hence, segregation analysis had to be replaced by mutation burden analysis as the main analytic approach. Because of the modest effect sizes and low allele frequencies, large sample sizes of tens to hundreds of thousands of individuals became necessary for disease-association studies. It was also apparent that the simple regression models used for testing of genetic–phenotype associations are underpowered for rare variants.[6] Specifically, the higher number of rare independent variants compared with common variants dramatically increases the requirement for multiple testing corrections. To increase the statistical power, combined information from multiple rare variants within a gene is often used. These approaches are grouped in 2 main categories: (1) the burden test that collapses genetic variants into a single score, assuming that tested variants have the same direction and magnitude of effect. This approach ignores the possibility that certain variants in the same gene may be neutral or have opposite effects. (2) Variance-component test that allows for different directions of effect, that is, risk and protective alleles. Unfortunately, this and many offshoots of this analytic tool are all far from perfection and have shown major practical limitations. In fact, most success in this venue has come from next-generation sequencing in case–control association studies of rare variants in candidate genes and genes in GWAS loci.[7] Not unexpectedly, the limitations of the analytic techniques have led to widespread use of lenient criteria in genetic studies and subsequent generation of false-positive results. Many genetic variants identified in single cases and small-size studies have been reported as independently disease causing without use of stringent criteria.

De novo mutations represent the most unique form of rare genetic variation because of their extremely low incidence. These low hanging fruits have been practically considered as pathogenic, in both small family-based and large case–control trio studies. In genetic classification guidelines of the American College of Medical Genetics and Genomics, these variants have been considered strong supporting evidence for pathogenicity.[1] Identification of these variants has been of great interest for genetic studies of severe traits such as congenital heart disease (CHD): a complex trait with few known genes. Trio studies, which use genetic data from case and parents, have indeed provided important insight into the pathogenicity of various CHD.[8] As estimated by these studies, the identified de novo variants account for ≈10% of severe CHD. The prevalence of damaging de novo variants in highly heart expressed genes have been estimated to be as high as 20% and 2% for syndromic and isolated CHD, respectively.[9] One key unanswered question is whether these variants are independently disease causing or only contributing to the disease. De novo mutations do not explain the recurrence of the disease in families, and consequently, their causality cannot be verified by segregation analysis. Rare variants can be fixed in certain populations because of well-known bottle-neck effect or genetic drift. Such limitation can give rise to false-positive results, especially when case–control populations are not ethnically matched or are small in size. Particularly, the presence of numerous disease-associated de novo variants in variant databases of the general population has generated doubt about their pathogenicity.[10] Consequently, the true causality of many previously identified disease-associated variants has been recently questioned.[11] These findings underscore the need for use of large control cohorts when studying genetic basis of common diseases.

In their article, Paludan-Müller et al[12] examined the pathogenicity of published de novo variants associated with severe arrhythmias and structural heart diseases by comparing their allele frequencies in the reference Exome Aggregation Consortium (ExAC) database with the expected prevalence for the associated diseases. The goal was to examine whether they are so-called standing variations in the general population. Because ExAC database consists of subjects free of severe disease, the presence of disease-associated variants in this database would exclude low reproductive fitness and question their pathogenicity. The authors studied 396 articles reporting genetic mutations for all syndromic and nonsyndromic cardiomyopathies, malignant ventricular arrhythmias, and CHDs. The study population included monogenic variants identified in isolated single cases and small studies (n<200, group A) and variants that increase the susceptibility for CHD, identified in 3 large cases and controls trio studies (n>1000, group B). De novo variants that are observed as standing or recurrent variations in ExAC were referred to as class 2 and otherwise as class 1 de novo variants. In group A, 211 de novo variants were identified with 11% categorized as class 2 variants. The total allele count in ExAC at class 2 sites was 109 in ≈844 theoretically expected cases, which would explain 13% of the disease burden. Strikingly, the genetic variants for Brugada syndrome had an extreme recurrence rate of ≈50%, with 155% of expected Brugada cases in ExAC being caused by only 4 variants. In addition, 1 variant would explain 29% of dilated cardiomyopathy cases and 4 variants would account for 23% long-QT syndrome cases. De novo variants identified as CHD causing had also high recurrence rate of ≈10% in a database that has low prevalence of CHD. These findings contradict the results from most earlier genetic studies and question the causality of many rare variants identified in single cases and small studies.

Twenty-six percent of de novo and 18% of chromatin-modifying variants in group B were present in the ExAC database. Overall, large proportion of variants for heart diseases would be accounted for by only 21 sites in the exome if considered as truly monogenetic. Overall conclusion of the study is that the variants in the ExAC database are misclassified as highly penetrant pathogenic de novo variants for cardiovascular diseases. As the authors also cautiously stated, large trio studies aim to demonstrate disease susceptibility and not causality of rare variants. In addition, many of them use large reference databases[8] and hence are prone to fewer false-positive results compared with small studies. As the authors state, the larger the reference databases of human genetic variants, the higher the possibility of finding standing variants. ExAC represents a large database of 60 000 exomes from subjects who participated in various disease-specific and population genetic studies. The data are largely limited to protein-coding variants within the genome and contains an average of 1 variant in every 8 bases of the exome. It provides a valuable resource for the clinical interpretation of variants observed in patients with

rare diseases. It should be noted, however, that the absence of a rare variant even in this large database of ≈60 000 can be used only as suggestive and not proof for causality of a rare variant. It will not be surprising if in larger reference databases some so-called pathogenic variants will be discovered as standing variants. This is particularly true for variants identified in ethnic groups not represented in this database. The bottle-neck effect and genetic drift are potential sources of error for false-positive discoveries. In conclusion, a key step in any pipeline for the discovery of causal rare variants is the use of large reference databases that represent diverse ethnic groups.

Disclosures

Arya Mani is supported by R35HL135767 grant from NHLBI and Yale Liver Center Pilot Grant from NIDDK..

References

1. Kimber CH, Doney AS, Pearson ER, McCarthy MI, Hattersley AT, Leese GP, et al. TCF7L2 in the Go-DARTS study: evidence for a gene dose effect on both diabetes susceptibility and control of glucose levels. *Diabetologia*. 2007;50:1186–1191. doi: 10.1007/s00125-007-0661-9.
2. Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet*. 2013;14:661–673. doi: 10.1038/nrg3502.
3. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al; 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68–74. doi: 10.1038/nature15393.
4. Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*. 2016;354:aaf6814. doi: 10.1126/science.aaf6814.
5. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*. 2012;336:740–743. doi: 10.1126/science.1217283.
6. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al; GoT2D Consortium. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet*. 2015;11:e1005165. doi: 10.1371/journal.pgen.1005165.
7. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011;43:1066–1073. doi: 10.1038/ng.952.
8. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. *Nature*. 2013;498:220–223. doi: 10.1038/nature12141.
9. Homsy J, Zaidi S, Shen Y, Ware JS, Samocha KE, Karczewski KJ, et al. De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science*. 2015;350:1262–1266. doi: 10.1126/science.aac9396.
10. Kosmicki JA, Samocha KE, Howrigan DP, Sanders SJ, Slowikowski K, Lek M, et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet*. 2017;49:504–510. doi: 10.1038/ng.3789.
11. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med*. 2016;375:655–665. doi: 10.1056/NEJMsa1507092.
12. Paludan-Müller C, Ahlberg G, Ghouse J, Svendsen JH, Haunsø S, Olesen MS. Analysis of 60 706 exomes questions the role of de novo variants previously implicated in cardiac disease. *Circ Cardiovasc Genet*. 2017;10:e001878. doi: 10.1161/CIRCGENETICS.117.001878.

KEY WORDS: Editorials ■ arrhythmias, cardiac ■ exome ■ genetic variation ■ heart disease, congenital

Pathogenicity of De Novo Rare Variants: Challenges and Opportunities

Arya Mani

Circ Cardiovasc Genet. 2017;10:
doi: 10.1161/CIRCGENETICS.117.002013
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1942-325X. Online ISSN: 1942-3268