The Case of “Missing Causal Genes” and the Practice of Medicine

A Sherlock Holmes Approach of Deductive Reasoning

Ali J. Marian

The practice of medicine is an exceedingly gratifying privilege, a notion that is commonly shared among the physicians. Also commonly understood but often not expressed is the humbling nature of this practice. Seldom, if ever, we know the true answers to our patients’ questions. Yet, we adeptly offer explanations, commonly based on the group data, and yet the essence of the practice of medicine is recognizing the uniqueness of the individual and his/her disease. To quote Sir William Osler, the father of western medicine, “the good physician treats the disease; the great physician treats the patient who has the disease.” Still, our answers typically gratify the patients and perhaps give the physicians an incontrovertibly false sense of being omniscient.

The landscape of the practice of medicine has been recently enriched with the entry of the medical genetics. The impetus provided by the recent elucidation of the genetic causes of single-gene cardiovascular diseases has ushered in clinical applications of the molecular genetic discoveries and along with it our exuberance about genetic-based diagnosis, risk stratification, prevention, and treatment. Molecular genetic testing is now routinely performed as part of the medical evaluation of patients with medical conditions commonly recognized as single-gene diseases, such as inherited cardiomyopathies and arrhythmic syndromes.1–3 It is also being advocated for risk stratification of patients with complex phenotypes, albeit its validity remains unsettled. Along with these advances, the need to educate and train the practicing physicians in understanding the fundamentals of genetics and their applications to the practice of medicine are increasingly recognized.4

Despite the remarkable progress in elucidation of genetic causes of cardiovascular diseases during the past 25 years, the causal genes in ≈30% to 50% of various single-gene cardiovascular diseases, including hereditary cardiomyopathies and arrhythmic syndromes, have yet-to-be identified (Figure).

The common causal genes, typically mapped through robust genetic linkage studies in large families, are responsible for about half and at most two third of single-gene cardiovascular diseases. The remaining genes are uncommon and typically occur in the sporadic cases and small families, which are not suitable for genetic linkage studies. Consequently, the approach to identify the remaining elusive causal genes for single-gene cardiovascular disorders has shifted toward DNA sequencing, primarily through whole-exome sequencing (WES) and less commonly, whole-genome sequencing.

The findings from whole-genome sequencing and WES projects have illustrated the complexity of the human genome. Fundamental to this complexity is the rare error rate of the DNA replication machinery, which is empirically calculated at about 1.0×10−8 per nucleotide.5,6 Consequently, each newborn (or meiosis) introduces ≈50 to 60 de novo genetic variations to the existing pool. The explosive growth of the human population during the recent epochs has introduced a huge number of genetic variants in the population, rendering the humans genetically extremely diverse. Recognition of such genetic complexity has infused a dose of sobriety as well as a sense of intimidation, because of the challenges it has posed in applying the discoveries to the practice of medicine. Genetic testing laboratories have taken the conservative approach in reporting only the well-characterized genes and variants, an appropriate measure to avoid overinterpretation, particularly by those lacking an adequate training in clinical genetics. And yet this approach often leads to a null test result and is anticlimactic. It is often a source of a disappointment for the patients and the physicians alike. The disappointment is even greater for the physician–scientists who are expected not only to make the medical discoveries and pave the way for tomorrow’s practice of medicine but also apply the new discoveries to the care of their patients.

Staring into the read out of the exome sequencing of a proband with hypertrophic cardiomyopathy (HCM), whose genetic pathogenesis is among the best characterized cardiovascular diseases,7 the daunting task is to identify the causal variants among >20,000 heterozygous variants that reside in the exons and exon–intron boundaries. WES approach also captures a huge number of variants in regions adjacent to the exons, which do not code for proteins. A Sherlock Holmes approach of deductive (or inductive) reasoning is needed to deduce the putative causal variant(s). Considering that the vast majority of the known causal mutations for single-gene disorders, including HCM, are located in the protein-coding regions (http://www.hgmd.org), it is logical to follow the bank robber Willie Sutton principle (“because that is where the money is”) and focus on the exons. The exon-centric approach...
has the inherent risk of the “winner’s curse,” and eliminating the chance of discovering new genetic mechanisms. An alternative approach of whole-genome sequencing that also encompasses introns and intergenic regions, however, adds further complexity, because of the presence of $\approx 4$ million genetic variants in each genome, the vast majority of which have yet to be annotated. $^{8, 9}$ Whether the current exome-centric approach is responsible, in part, for failure to identify the “missing causal genes” for cardiovascular single-gene disorders remains an open question.

Having settled on following the Willie Sutton principle, at the risk of the “winner’s curse,” the focus is on the variants that affect structure and potential function of their respective encoded proteins. The approach eliminates the synonymous variants, even though not all synonymous variants are innocuous, as best illustrated for the synonymous p.G608G variant in the LMNA gene, which activate a cryptic splice site, and leads to Hutchinson–Gilford progeria syndrome. $^{10}$ Nevertheless, despite this generally accepted but not flawless filtering approach, the remaining list includes $\approx 10,000$ nonsynonymous, 150 frameshift, 350 nonframeshift insertion/deletion, and 100 stop codon gain or loss variants, all having the potential to be functional and pathogenic. Computational algorithms that incorporate population variant frequency, evolutionary conservation of the involved amino acids, and the predicted effects of the variant on biophysical and biochemical properties of the involved protein are useful in discerning the pathogenic and deleterious variants. Regardless of the prediction by in silico algorithms, however, it is hard to argue that $>600$ variants that are expected to affect their respective protein length and structure are biologically inert and innocuous. Eliminating variants that are more common in the population than the prevalence of HCM, estimated at $\approx 1$ in 500 in the general population of young adults, $^{11}$ and considering genetically heterogeneity of HCM, $^{7}$ one could further narrow down the search to rare putatively functional variants. The filtering approach and integration of the biological data, further restricts the candidate causal variants to a handful. In best situations, one of a rare pathogenic candidate variant resides in a gene known to cause HCM, and hence, is isolated, with a reasonable accuracy, but not with an unambiguous certainty, as the causal variant/gene. Otherwise, the approach has met the “Peter Principle,” rendering identification of the causal variant challenging, if not impossible, even if the candidate variant resides in a biologically plausible candidate protein. Moreover, the failure to identify might also reflect, despite all technical advances, technical imperfectness of WES variant detection, annotation, and filtering algorithms.

To narrow down the list of the candidate causal variants in a proband or members of a small family, a commonly used approach, typically for molecular diagnostic purposes, is to perform WES followed by a focused analysis restricted to genes known to cause the phenotype of interest. This targeted

---

**Nonstandard Abbreviations and Acronyms**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCM</td>
<td>hypertrophic cardiomyopathy</td>
</tr>
<tr>
<td>WES</td>
<td>whole-exome sequencing</td>
</tr>
</tbody>
</table>

---

Figure. **Known and yet-to-be identified causal genes for hypertrophic cardiomyopathy (HCM).** The common causal genes, identified through linkage analysis in multiple large families, are responsible for $\approx 60\%$ of HCM. Evidence for causality for such genes is strong. In contrast, the causal genes in $\approx 40\%$ of HCM, occurring in small families and sporadic cases, remain unknown. Causality is difficult to establish in such cases.
approach typically entails analysis of ≈50 to 100 known causal genes, as opposed to ≈20000 genes in the human genome, and therefore, irrefutably, reduces the number of the putative candidate pathogenic variants. However, this focused approach explicity is not applicable to discovering novel genes and inherently has a high rate of a negative test result, as the causal gene(s) in ≈30% to 50% of each of the single-gene cardiovascular disease have yet-to-be-identified. Moreover, it merits to be noted that “no gene or protein is a perfect molecule.” Accordingly, even the best-established causal genes for monogenic cardiovascular diseases carry variants that not only are considered functional but also pathogenic, as predicted by various in silico algorithms, and yet they occur in individuals free of the relevant clinical phenotype. A notable example is the case of TTN, encoding the giant sarcomere protein titin, which is the most common causal gene for dilated cardiomyopathy.12 Yet, the gene carries a large number of missense, stop codon, and frame shift (truncating) variants that occur in normal individuals.12–14 In the best-case scenario in the candidate gene approach, the list of the pathogenic variants might include a previously established causal variant. Such finding typically satisfies the inquisitive physician for discovering a genetic pathogenesis for his/her patient’s condition. Yet, the extrapolation dismisses the possibility of the potential effects of the genetic background and the environmental factors in precipitating the causality in one specific individual but not in the other and hence, restricting generalizability of the previous findings.

Given that establishing causality of a genetic variant in a single individual unambiguously, even with a well-defined single-gene disorder, is almost impossible, it is best to focus on families, as inclusion of each family member could enable reducing the number of the candidate genes/variants by ≤50%. The causal genes in most large families with single-gene disorders have already been identified (Figure). Therefore, one is often limited to analyzing the data in nuclear families composed of often 2, sometimes 3 individuals, and occasionally more. At such situation, one, at best, might infer causality, but seldom is able establish it with a high level of confidence.

The discoveries ushered in by the large-scale DNA sequencing technologies are also blurring the classic categorical distinction between single-gene and complex disorders, rendering the disease spectrum a continuum, wherein the classical single-gene and complex disorders represent the opposite ends of the spectrum.15,16 Accordingly, a subset of the “single-gene disorders” might be digenic and even oligogenic, wherein ≥2 distinct pathogenic variants are responsible for the phenotype, whereas none perfectly cosegregating with inheritance of the phenotype. An oligogenic pathogenesis might explain the elusive genetic pathogenesis of a subset of the conventionally considered single-gene disorders. However, it is practically impossible to ascertain causality of a single, let alone multiple variants, in a single individual and even in members of a small family, in an unambiguous manner. Thus, to test an oligogenic pathogenesis of a subset of presumably single-gene disorders, therefore, one has to depend on the group data and apply various statistical methods to assess enrichment of the pathogenic variants in the genes of interest in those with the phenotype. Statistical evidence typically mandate complementation with in vitro and in vitro functional studies to fulfill the analogous elements of the Koch postulates of causality.17,18 Until then, whether such digenic or oligogenic pathogenesis could, in part, explain the “missing genes” in ≈30% to 50% of the conventionally considered single-gene cardiovascular disorders is an open question in need of empirical testing. Such pathogenic complexity along with possible involvement of copy number variants and mitochondrial DNA have posed further challenges in the clinical implications of genetic screening.

The view presented probably reflects the conviction that the most profound impact of molecular genetic discoveries is in elucidating the fundamental mechanisms that govern the pathogenesis of the disease. This notion is best illustrated in the case of PCSK9 gene, which was mapped in a family with a rare form of autosomal-dominant hypercholesterolemia.19 Within a decade or so, the discovery led to development of powerful new therapies for patients with hypercholesterolemia.20–22 Thus, despite recent technological advances and our exuberance in application of the modern genetic discoveries to the practice of medicine, the complexity of the human molecular genetics has eluded us in identifying the remaining “missing causal genes” in patients/families with single-gene cardiovascular disorders. The complexity of the human genome has made it amply evident that “the more we know, the more we know that we do not know.”

Sources of Funding

This study was supported, in part, by grants from National Institutes of Health, National Heart, Lung and Blood Institute (R01 HL088498, 1R01HL132401, and R34 HL105563), Leducq Foundation (14 CVD 03), Roderick MacDonald Foundation (13RDM005), TexGen Fund from Greater Houston Community Foundation, Texas Heart Institute Foundation, and George and Mary Josephine Hamman Foundation.

Disclosures

None.

References


**Key Words:** genetic testing ■ hypertrophic ■ introns ■ meiosis ■ mutation ■ variant
The Case of "Missing Causal Genes" and the Practice of Medicine: A Sherlock Holmes Approach of Deductive Reasoning
Ali J. Marian

*Circ Res.* 2016;119:21-24
doi: 10.1161/CIRCRESAHA.116.308830

*Circulation Research* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2016 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circres.ahajournals.org/content/119/1/21

**Permissions:** Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation Research* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the *Permissions and Rights Question and Answer* document.

**Reprints:** Information about reprints can be found online at:
http://www.lww.com/reprints

**Subscriptions:** Information about subscribing to *Circulation Research* is online at:
http://circres.ahajournals.org/subscriptions/