



Published in final edited form as:

*Circ Cardiovasc Genet.* 2009 April ; 2(2): 95–97. doi:10.1161/CIRCGENETICS.109.864793.

## Defining the pathogenicity of DNA sequence variation

Carolyn Y. Ho<sup>1</sup> and Calum A. MacRae<sup>2</sup>

<sup>1</sup>Cardiovascular Division Brigham and Women's Hospital 75 Francis Street Boston MA 02115

<sup>2</sup>Cardiovascular Research Center and Cardiology Division Massachusetts General Hospital 55 Fruit Street Boston, MA 02114

---

Understanding the relationship between genotype and phenotype remains one of the most challenging hurdles in human genetics, especially as efforts are made to translate genetic data into clinical prediction of disease or therapeutic outcomes[1].

### Mendelian disorders

For many cardiac or vascular conditions where the initial presentation may be fatal, a strong genotype-phenotype correlation is of fundamental importance if genetic diagnosis or prognostication is ever to be of utility. Even in classic monogenic disorders, where large effect sizes are observed, genetic prediction is often confounded by reduced penetrance, wide variation in the final phenotypes (pleiotropy), or sporadic phenocopies[2]. While such trait 'complexity' has been viewed largely in terms of environmental or genetic modifiers, such factors have proven difficult to identify. It is also possible that strong selection pressures against many cardiovascular traits may result in purely stochastic events playing a larger role in shaping phenotype[3]. Rigorous genetic studies have identified many cardiovascular disease genes, leading to broader investigation in smaller families and individual probands. Although identifying variants in these same genes in unrelated individuals provides important confirmation and strengthens the initial data, the clinical significance of the numerous additional variants uncovered in subsequent re-sequencing efforts may be ambiguous[4]. The level of support that these variants can cause disease is often much less robust than in the original discovery of causal genes. As a result rare (or not so rare) benign polymorphisms populate many mutation databases, often indistinguishable from pathogenic mutations.

The interpretation of sequence data in the absence of relevant genetic or functional data to support pathogenicity is fraught with problems. The pertinent issues are well illustrated in research and commercial proband-based "gene screening" efforts in several cardiac syndromes where detected rates of *de novo* DNA sequence variants are high, reliable data to predict their clinical significance are scarce, long-term outcomes are typically not available, and single individuals are occasionally found to have mutations in multiple genes[5]. In these latter cases, multiple mutations are often proposed to result in more profound phenotypes due to a putative gene dosage effect. Unfortunately, it can be extremely challenging to determine the functional consequences of all sequence variants since the background phenotypic heterogeneity is substantial. Furthermore, family segregation data, detailed clinical outcomes, or functional assays are often not available for one, let alone two,

---

617-726-4343 617-726-5806 macrae@cvrc.mgh.harvard.edu cmacrae@partners.org.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

variants. Rigorous evidence to support the true impact of double mutations is available for remarkably few reported cases making it difficult to accurately speculate on the clinical significance of multiple mutations on either an individual or a population basis[1,5].

The absence of robust genetic support in humans or in animal models is further compounded by reliance on other circumstantial lines of evidence. Re-sequencing studies of disease genes have typically not subjected control populations to the same level of scrutiny as patient cohorts, therefore similar rare but benign variants in normal individuals go undetected. Even if available, *in vitro* functional data obtained in heterologous systems may not accurately reflect *in vivo* physiology as a result of failure to recapitulate key components of the cellular or organismal biology[6]. Partner proteins, intracellular compartmentalization of signals, unknown environmental contributions, as well as significant pathway redundancy all may modify the final phenotype[3]. Nucleotide variants with apparent phenotypes *in vitro* may have no effect in the context of the powerful homeostatic influences of other physiologic pathways. Human genetics offers many examples of profound *in vitro* effects that fail to translate into an obvious *in vivo* phenotype[7]. As an end result, pathogenicity may be attributed to DNA sequence variants using flawed and somewhat circular logic.

### Common disease: the challenges of GWAS

These challenges for genotype-phenotype correlation are only magnified in common traits, where underlying etiologic heterogeneity may be considerable, heritability is less clearly defined, and effect sizes are orders of magnitude smaller[8]. These very factors will also militate against rapid mechanistic understanding of the role of individual variants. Detection of subtle coding sequence alleles, non-coding regulatory functions or remote effects in *cis* and *trans*, are all likely to be compromised by intrinsically low signal:noise ratios[8–10]. Given the limited predictive utility of genetic data in many Mendelian diseases, it is difficult to imagine at present how individualized prediction at the genomic level will be feasible with current genetic datasets[1]. The cumulative risk attributable to the major common alleles for any given trait is unlikely to translate into meaningful insights if, as has been the case in most studies to date, molecular mechanisms and larger genetic or environmental factors remain uncharacterized. Any attempt to define robust genotype-phenotype correlations must address these unknown intermediate effect alleles, many of which might have been predicted from preexisting heritability data.

### Potential Solutions

Ultimately, several different strategies must converge to facilitate the development of quantitative relationships between nucleotide variants and clinically meaningful traits. As they emerge, massively parallel sequencing technologies will allow large-scale re-sequencing studies in patients and controls to accelerate the identification of genetic associations with disease[11]. Nevertheless, these studies will be subject to many of the confounders noted above, particularly if rooted in heterogeneous cohorts with imprecise phenotypes.

There is no substitute for a detailed understanding of the mechanisms of disease. Disease-relevant functional assays have proven useful in developing predictive frameworks especially where a biologically relevant activity of the protein in health and disease can be measured directly *in vitro* or *in vivo*. Yet even in deficiencies of a single enzyme, correlation between a specific *in vitro* assay and clinical severity can be difficult to capture[12]. Where less is known of the causal chain between DNA sequence variation and final phenotype, or where this relationship is difficult to discriminate above the biological noise, any such correlation falls off rapidly. For example, disease may result from subtle effects on development decades earlier, the protein may have unsuspected attributes, or mutation may

lead to gain of novel functions. It is no surprise that these aspects are not often captured by conventional clinical phenotypes that typically reflect only macroscopic late-stage features of disease. Pleiotropy and reduced penetrance, both commonly seen in heritable conditions, challenge efforts to garner useful predictive genotypic data. Experience with monogenic disorders as well as emerging results from genome wide association studies suggest that the apparent 'complexity' of common disease often resides in the phenotype[3,13].

Examples of this 'complexity' abound. In cardiac repolarization, despite the mechanistic insights from the successful cloning of the Long QT genes, predicting sudden death from simple genotype has proven challenging even within a single family. For many of the sequence variants identified to date, there are few genetic data to substantiate a causal role in arrhythmia. The QT interval varies widely with autonomic and environmental influences, fails to encompass subjective morphologic abnormalities of repolarization and is at best an imperfect reflection of the underlying proarrhythmic state[14,15]. There is often discordance between the available *in vitro* data from heterologous expression studies of ion channel function and observed effects on ventricular electrophysiology in animal models or in man[14,16]. Few if any studies correlating clinical outcome with genotype account for the relatedness of individuals, and apparent risk factors for sudden death are thus more likely to reflect attributes of specific families within the cohorts than provide robust measures of the utility of genotype.

The next generation of patient-oriented studies must offer rigorous estimates of heritability and more precise definition of disease traits. Currently such efforts are constrained by the limited resolution of traditional clinical assays[13,17]. Identifying more granular and quantitative novel disease traits or endophenotypes will not only help to resolve etiologic heterogeneity (even if acquired), but also will enable empiric estimates of the genetic architecture of these traits, improve the power of all forms of genetic study, and allow the cost effective prioritization of investigation based on magnitude of effect. These new clinical studies will exploit the latest functional genomic, proteomic and metabolomic technologies, as well as reappraise traditional phenotypes ranging from biochemistry and physiology to imaging[18,19]. Perhaps the most effective strategies will integrate multiple approaches at a systems level, building disease specific networks that may then be interrogated[18,20,21]. As we understand the complete genetic architecture of each disease, so the relative importance of specific genotypes will be more readily estimated and robust population attributable risks may be calculated[3]. Validated endophenotypes will offer immediate benefits to all forms of genetic study, expanding Mendelian families and resolving heterogeneity in cohorts for genome-wide association[8,13].

Ultimately, in the face of a need for both relevant functional assays to test the effects of genotypic variation, and more multidimensional clinical phenotypes, it may be most efficient to focus efforts on the identification of endpoints that will serve both purposes. Clinical relevance will be key, and the most useful phenotypes are likely to encompass dynamic responses and allow bidirectional translation from bench to bedside. For most diseases, this will require systematic approaches to redefine the most informative components of the underlying pathologic traits[18]. This 'phenome' project will benefit from the investment in post-genomic technologies but must leverage the legacy of classic clinical investigation[13]. Dynamic responses to small molecule probes (drugs or diagnostics) are one way in which directed efforts at phenotype discovery might be undertaken. These efforts might occur *in vivo*, as seen in recent work using ion channel blockade to characterize repolarization reserve in individual subjects[15].

The power of combining biologically relevant assays with rich understanding of genetic mechanisms is already evident in malignant clonal disorders, where genotypic prediction of

the response to therapy has rapidly been implemented in the clinical arena[22]. Overcoming the challenges to accurate definition of the pathogenicity of sequence variants is a major hurdle in Mendelian disease, but critical for realizing the full potential of genomic medicine. Developing more sophisticated methods to redefine disease phenotypes will be an important step in what may be a longer but exciting cycle of discovery and translation in the cardiovascular field. Ultimately, this new phenome may complement ongoing efforts to redefine the healthcare delivery interface.

## References

1. Janssens AC, et al. A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet.* 2008; 82(3):593–9. [PubMed: 18319070]
2. Arad M, Seidman JG, Seidman CE. Phenotypic diversity in hypertrophic cardiomyopathy. *Hum Mol Genet.* 2002; 11(20):2499–506. [PubMed: 12351586]
3. Fraser AG, Marcotte EM. A probabilistic view of gene function. *Nat Genet.* 2004; 36(6):559–64. [PubMed: 15167932]
4. Ellinor PT, MacRae CA. Ion channel mutations in AF: signal or noise? *Heart Rhythm.* 2008; 5(3):436–7. [PubMed: 18313603]
5. Kelly, Semsarian. Multiple mutations in genetic cardiovascular disease: a marker of disease severity. *Circulation Cardiovascular Genetics.* 2009 In press.
6. Remme CA, Wilde AA, Bezzina CR. Cardiac sodium channel overlap syndromes: different faces of SCN5A mutations. *Trends Cardiovasc Med.* 2008; 18(3):78–87. [PubMed: 18436145]
7. North KN, et al. A common nonsense mutation results in alpha-actinin-3 deficiency in the general population. *Nat Genet.* 1999; 21(4):353–4. [PubMed: 10192379]
8. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet.* 2008; 17(R2):R156–65. [PubMed: 18852205]
9. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature.* 2007; 447(7146):799–816. [PubMed: 17571346]
10. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet.* 2001; 2(2):91–9. [PubMed: 11253062]
11. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008; 26(10):1135–45. [PubMed: 18846087]
12. Aerts JM, et al. Elevated globotriaosylsphingosine is a hallmark of Fabry disease. *Proc Natl Acad Sci U S A.* 2008; 105(8):2812–7. [PubMed: 18287059]
13. Anttila V, et al. Trait components provide tools to dissect the genetic susceptibility of migraine. *Am J Hum Genet.* 2006; 79(1):85–99. [PubMed: 16773568]
14. Roden DM. Defective ion channel function in the long QT syndrome: multiple unexpected mechanisms. *J Mol Cell Cardiol.* 2001; 33(2):185–7. [PubMed: 11162124]
15. Roden DM. Drug-induced prolongation of the QT interval. *N Engl J Med.* 2004; 350(10):1013–22. [PubMed: 14999113]
16. Casimiro MC, et al. Targeted point mutagenesis of mouse *Kcnq1*: phenotypic analysis of mice with point mutations that cause Romano-Ward syndrome in humans. *Genomics.* 2004; 84(3):555–64. [PubMed: 15498462]
17. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* 2008; 322(5903):881–8. [PubMed: 18988837]
18. Leboyer M, et al. Psychiatric genetics: search for phenotypes. *Trends Neurosci.* 1998; 21(3):102–5. [PubMed: 9530915]
19. Singer E. “Phenome” project set to pin down subgroups of autism. *Nat Med.* 2005; 11(6):583. [PubMed: 15937456]
20. Walhout AJ, et al. Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr Biol.* 2002; 12(22):1952–8. [PubMed: 12445390]

21. Lee I, et al. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat Genet.* 2008; 40(2):181–8. [PubMed: 18223650]
22. Lynch TJ, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med.* 2004; 350(21):2129–39. [PubMed: 15118073]