Genome Biology

COMMENT

# Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases

Isaac S Kohane

A recent National Academy of Sciences report entitled 'Precision Medicine' [1] made the point that, in this era of commodity-priced genome-scale measurements, we can now envisage a systematic reclassification of human pathobiology on a population scale. These high-throughput measurement modalities promise greater precision and accuracy to provide patients with individualized diagnoses and therapies. Indeed, we have already seen remarkable success in this regard in improved prognostics and therapeutics for breast cancer [2], non-small-cell lung carcinomas [3] and the leukemias [4] through molecular-based subtype profiling. By contrast, many have written about the artificiality of current organ-based phenotypes and often clinical-department-based diagnoses [5,6] that do not correspond to the underlying pathotypes that cross conventional clinical categorizations. This inadequacy of the current and often-arbitrary clinical classifications, coupled with encouraging results from molecular medicine, has led to a swing of the pendulum to the opposite extreme of where it was in the pre-genomic era. Genotypic variation is often but a small slice of relevant pathotypic variation [7], and the recent call for a sequencing-first approach [8] for molecular-driven classification could result in expensive and frustrating delays in discovering the true genetic architecture of much of human disease. In many cases, taking a more detailed data-driven look at the clinical characterization of individual patients, particularly as revealed by their distinct trajectories over time, might rescue a large number of otherwise-misdirected genomic investigations.

Premature categorization of a clinical phenotype in a genomic case–control study, particularly in complex disease, can lead to an injudicious investment of limited resources for a restricted scientific payoff. First, for example, consider a reasonably common disease, such as autism, affecting over 1% of individuals. Suppose that, like many common diseases, it is suspected that its inherited component is caused by a large set of genetic sequence variants in different genes and even different pathways [9-13]. If each of the disease-causing variants is even modestly rare, then a simple case-controlled study will require numbers of patients orders of magnitude higher than the investigators might be able to recruit. For example, if the disease prevalence is as high as 1%, variant frequency of 1%, relative risk of 2.0, then, with 80% power, discovering each of these variants would require 23,000 subjects [14], which will typically take many years and cost millions of dollars. This imposes a delay to the time when we can better understand the genetic architecture of the disease. Second, it presents a significant economic burden in times of difficult funding for science. This problem is accentuated by the inevitable contributions of noise and bias of environmental exposures to the phenotypic variance as well as the gene-environment interactions [15].

However, with a phenotypic-driven longitudinal approach, researchers will observe individuals who develop clinical findings that are not the primary disease phenotype but are instrumental in understanding the pathobiology of the patient. If the additional clinical findings (for example, co-morbidities) are themselves uncommon (for example, found in 2% or less of the individuals with the primary disease phenotype), even the clinicians caring for the patients might not recognize that there exist groups of patients with archetypal clusters of these co-morbidities. There will therefore be subpopulations of clustered clinical pathologies that would be completely opaque to the original classic approach for a genomic association study. A new and potentially more powerful paradigm for genomic association would include identification of the genetic architectures behind each phenotypic cluster. If there are, for example, 10 such (similarly sized) phenotypic clusters, the frequency of variants that contribute to phenotypes of individual clusters can increase and be as large as 10%. Similarly, they can drop to 0% for those clusters that they do not contribute to. In that case, with a relative risk of 2.0, only 2,300 subjects would have to be studied rather

Correspondence: isaac_kohane@harvard.edu
Center for Biomedical Informatics, Harvard Medical School, Boston, MA 01115, USA

than 23,000 - a change that might make the difference between a successful or a disappointing study. Of course, there are several limiting assumptions implicit in this scenario, including first that the individual genetic variants are contributing to the frequency of co-morbidities in each cluster and, second, that the contributions of the individual variants are identical for each of these co-morbidities.

Despite the aforementioned caveats, we have many examples where better phenotyping enables better understanding of the genetics of disease. For example, whereas 100 years ago heart failure was viewed as a monolithic disease, careful current phenotyping and population studies have revealed heart failure in middle-aged individuals who are highly enriched for the cardiomyopathy gene variants. Similarly, older individuals suffering heart failure due to atherosclerosis have a different set of variants that contribute to the disease. After the fact, it would seem ludicrous to perform a case–control study across all heart failure patients - but, in effect, that is how many of our current and planned studies are structured, although there are notable exceptions (for example, in diabetes [16] and asthma [17]).

As a research community, we can now break free from our definitions of disease and allow the full biological impact of the genetic variants to be expressed across time and across multiple symptom complexes. An important and previously definitive objection to this approach was simply one of cost. Whereas the cost of a whole-genome variant scan is $100 or less and a whole-genome sequence is $1,000, characterizing a patient fully and repeatedly over their lifetime can and will cost many tens of thousands of dollars. Fortunately, as a by-product of the automation of healthcare, there are increasingly large volumes of data that are available across years and decades of a patient's lifetime over which thousands of different clinical variables are measured [18,19]. Clinical narrative notes from the electronic health record can also be turned into codified variables through the process of natural language processing [20,21]. This now allows the identification of clusters of patients arising over time at a marginal cost of cents per patient and at very high speed. For example, in a recent study of children with autism, it was possible to identify clusters of children with autism and 80% prevalence of seizures, another subgroup with a high prevalence of viral and bacterial infections, and autoimmune diseases, and a third group with a variety of neuropsychiatric diseases such as schizophrenia, attention deficit hyperactivity disorder and anxiety disorders [22]. Therefore, rather than a monolithic disease, autism begins to look more like a set of clinical syndromes that each merits its own independent genetic study, just like the distinct causes of heart failure.

It will become clearer over time that, in this instance, we can have our cake and eat it too. A deeper and longer phenotyping of human populations is more possible than ever before with emerging big-data sets, such as access to biorepositories and longitudinal troves of real-time health information on patients. Just as predicted in the precision-medicine report [1], we can create at minimal incremental cost an 'information commons' of a large, even national, population that resolves to the single individual the full array of molecular, genome-scale characterizations. Furthermore, this will permit a deep characterization of the clinical evolution of each of these patients over time so that, in a genuinely data-driven fashion, we can determine what are the true or natural biologically coherent subclasses, whether driven by genetic or environmental influences.

#### Competing interests

The author declares that he has no competing interests.

#### Acknowledgements

IK thanks the following colleagues for their thoughtful comments and suggestions: Arjun Manrai, Chirag Patel and Nathan Palmer.

Published: 20 May 2014

#### References

1. National Research Council: *Toward Precision Medicine : Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.* Washington, DC: The National Academies Press; 2011.
2. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF: **The prognostic role of a gene signature from tumorigenic breast-cancer cells.** *N Engl J Med* 2007, **356:**217–226.
3. Raponi M, Dossey L, Jatkoe T, Wu X, Chen G, Fan H, Beer DG: **MicroRNA classifiers for predicting prognosis of squamous cell lung cancer.** *Cancer Res* 2009, **69:**5776–5783.
4. Kang H, Chen IM, Wilson CS, Bedrick EJ, Harvey RC, Atlas SR, Devidas M, Mullighan CG, Wang X, Murphy M, Ar K, Wharton W, Borowitz MJ, Bowman WP, Bhojwani D, Carroll WL, Camitta BM, Reaman GH, Smith MA, Downing JR, Hunger SP, Willman CL: **Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia.** *Blood* 2010, **115:**1394–1405.
5. Butte AJ, Kohane IS: **Creation and implications of a phenome-genome network.** *Nat Biotechnol* 2006, **24:**55–62.
6. Loscalzo J, Kohane I, Barabasi AL: **Human disease classification in the postgenomic era: a complex systems approach to human pathobiology.** *Mol Syst Biol* 2007, **3:**124.
7. Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE: **The predictive capacity of personal genome sequencing.** *Sci Transl Med* 2012, **4:**133ra158.
8. Stessman HA, Bernier R, Eichler EE: **A genotype-first approach to defining the subtypes of a complex disease.** *Cell* 2014, **156:**872–877.
9. Frayling TM, Lindgren CM, Chevre JC, Menzel S, Wishart M, Benmezroua Y, Brown A, Evans JC, Rao PS, Dina C, Lecoeur C, Kanninen T, Almgren P, Bulman MP, Wang Y, Mills J, Wright-Pascoe R, Mahtani MM, Prisco F, Costa A, Cognet I, Hansen T, Pedersen O, Ellard S, Tuomi T, Groop LC, Froguel P, Hattersley AT, Vaxillaire M: **A genome-wide scan in families with maturity-onset diabetes of the young: evidence for further genetic heterogeneity.** *Diabetes* 2003, **52:**872–881.
10. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir HT, Johannsdottir H, Magnusson OT, Gudjonsson SA, Justesen JM, Harder MN, Jørgensen ME, Christensen C, Brandslund I, Sandbæk A, Lauritzen T, Vestergaard H, Linneberg A, Jørgensen T, Hansen T, Daneshpour MS, Fallah MS, Hreidarsson AB, Sigurdsson G, Azizi F, Benediktsson R, Masson G, Helgason A, Kong A, *et al*: **Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes.** *Nat Genet* 2014, **46:**294–298.

11. Betancur C: **Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting.** *Brain Res* 2011, **1380:**42–77.
12. Van Bokhoven H: **Genetic and epigenetic networks in intellectual disabilities.** *Annu Rev Genet* 2011, **45:**81–104.
13. Campbell MG, Kohane IS, Kong SW: **Pathway-based outlier method reveals heterogeneous genomic structure of autism in blood transcriptome.** *BMC Med Genomics* 2013, **6:**34.
14. Purcell S, Cherny SS, Sham PC: **Genetic power calculator: design of linkage and association genetic mapping studies of complex traits.** *Bioinformatics* 2003, **19:**149–150.
15. Deary IJ, Yang J, Davies G, Harris SE, Tenesa A, Liewald D, Luciano M, Lopez LM, Gow AJ, Corley J, Redmond P, Fox HC, Rowe SJ, Haggarty P, McNeill G, Goddard ME, Porteous DJ, Whalley LJ, Starr JM, Visscher PM: **Genetic contributions to stability and change in intelligence from childhood to old age.** *Nature* 2012, **482:**212–215.
16. Ingelsson E, Langenberg C, Hivert MF, Prokopenko I, Lyssenko V, Dupuis J, Mägi R, Sharp S, Jackson AU, Assimes TL, Shrader P, Knowles JW, Zethelius B, Abbasi FA, Bergman RN, Bergmann A, Berne C, Boehnke M, Bonnycastle LL, Bornstein SR, Buchanan TA, Bumpstead SJ, Böttcher Y, Chines P, Collins FS, Cooper CC, Dennison EM, Erdos MR, Ferrannini E, Fox CS: **Detailed physiologic characterization reveals diverse mechanisms for novel genetic loci regulating glucose and insulin metabolism in humans.** *Diabetes* 2010, **59:**1266–1275.
17. Howrylak JA, Fuglbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA: **Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications.** *J Allergy Clin Immunol* 2014, **133:**1289–1300.
18. Murphy S, Churchill S, Bry L, Chueh H, Weiss S, Lazarus R, Zeng Q, Dubey A, Gainer V, Mendis M, Glaser J, Kohane I: **Instrumenting the health care enterprise for discovery research in the genomic era.** *Genome Res* 2009, **19:**1675–1681.
19. Kohane IS: **Using electronic health records to drive discovery in disease genomics.** *Nat Rev Genet* 2011, **12:**417–428.
20. Ananthakrishnan AN, Cai T, Savova G, Cheng SC, Chen P, Perez RG, Gainer VS, Murphy SN, Szolovits P, Xia Z, Shaw S, Churchill S, Karlson EW, Kohane I, Plenge RM, Liao KP: **Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach.** *Inflamm Bowel Dis* 2013, **19:**1411–1420.
21. Meystre S, Haug PJ: **Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation.** *J Biomed Inform* 2006, **39:**589–599.
22. Doshi-Velez F, Ge Y, Kohane I: **Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis.** *Pediatrics* 2014, **133:**e54–e63.

doi:10.1186/gb4175
**Cite this article as:** Kohane: **Deeper, longer phenotyping to accelerate the discovery of the genetic architectures of diseases.** *Genome Biology* 2014 **15**:115.