OPINION

The Faustian bargain in data-driven discovery: lessons from medicine

Anurag Agrawal

Technological progress now permits hypothesis-free acquisition and interrogation of genome-scale data[1]. This has allowed many investigators to pursue a genetic understanding of diseases, unmindful of concerns that plagued yesterday's scientists, such as which genes (if any) were likely to be involved. Low-hanging fruits being already plucked, the search for statistically significant small effects continues through ever-increasing sample sizes. Despite the limitations of a medical background, close interactions with the genomics community have permitted me to come up to speed with this field. I increasingly find that the experiences of the medical community with our version of technologically driven 'hypothesis-free' medicine are relevant to understanding the current state and future directions in the field of genomics. Despite easy availability of large-scale serum biochemistry-based health profiles, composed of a multitude of markers with statistically significant associations with diseased states, the benefit derived from performing hundreds of tests in the absence of any question beyond 'is there anything abnormal?', is limited. Many of these markers come from large, well-designed studies that succeeded in showing statistically significant but modest associations with increased risk of disease, but what exactly they mean and what to do with them is still unclear after decades[2]. It is known that such abnormalities usually lead to further studies and occasionally diagnoses and treatments, but more often than not remain just that – an abnormal value. This concept of 'executive physicals' or 'master health checkups' is now labelled bad medicine because of the potential for harm outweighing the benefit[3]. Yet, the entire process of obtaining even large panels of clinical biochemistry tests is extremely meaningful when done in the contexts of at-risk populations by experienced clinicians. The difference is Bayesian, with positive tests with low prior odds being less meaningful than a similarly positive test with high prior odds[4]. It takes medical knowledge to properly judge the risk before ordering the tests.

Genomics is no different. Poorly done studies can confuse more than clarify. Even in conceptually and technically correct studies where false positives are reduced by statistical treatment (at the cost of increasing false negatives), it still takes prior knowledge to prioritize results. A statistically significant hit in a highly relevant pathway with high prior odds is more meaningful and has a lower false positive rate probability, than a hit in a less relevant pathway with the same *P*-value[5]. What constitutes relevance is both knowledge and bias. Sometimes, as in medicine, the unexpected is true with large payoffs in discovering novel biology, and exploration of what is unknown has yielded many important results. Pursuit of such leads is reasonable when the odds are large and strongly against chance, analogous to finding a big mass on a screening chest X-ray; but very different from a small granuloma or a weak genetic association that may be statistically significant, but is associated with small risks. Whereas genetic findings associated with only small increase in risks are generally unsuitable for clinical use[6], those pointing in interesting or novel directions may at times be reasonable, especially when conventional thinking has failed. Yet, more often than not, the old medical adage 'When you hear hoof beats, think horses not zebras' remains true in science as well. Results from large-scale studies will often recapitulate what was known. Yet, the agreement of hypothesis-free approaches with the hypothesis-driven approaches is valuable, when it happens, because it reflects correct understanding and may then spur further research[7].

I do not intend to detract from the power of such approaches in contributing to understanding pathophysiology and phenotypes. Rather than perennially extending long-studied pathways, researchers can explore new genes found to have relevance to human health and disease. I only believe that while important discoveries will emerge from large-scale genomics (or any other omics), they will come from knowledgeable investigators with depth in molecular understanding of disease. Whether this knowledge should be called bias is debatable, but I contend that an unbiased mathematician with no knowledge of biology would do a poorer job. Until the new era of Bayesian statistics arrives, incorporating prior knowledge into statistics, physicians and investigators who understand health and disease can rest secure in the knowledge that no substitute is in sight.

1. Hunter, D. J., Altshuler, D. and Rader, D. J., *N. Engl. J. Med.*, 2008, **358**(26), 2760–2763.
2. Hoffman, R. M., *Curr. Opin. Urol.*, 2010, **20**(3), 189–193.
3. Rank, B., *N. Engl. J. Med.*, 2008, **359**(14), 1424–1425.
4. Goodman, S. N., *Ann. Intern. Med.*, 1999, **130**(12), 1005–1013.
5. Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L. and Rothman, N., *J. Natl. Cancer Inst.*, 2004, **96**(6), 434–442.
6. Agrawal, A., *Natl. Med. J. India*, 2009, **22**(3), 113–115.
7. Wang, Z. *et al.*, *Nature*, 2011, **472**(7341), 57–63.

Anurag Agrawal is in the CSIR Institute of Genomics and Integrative Biology, Mall Road, Delhi University Campus, Delhi 110 007, India.
e-mail: a.agrawal@igib.in