

BY ISAAC KOHANE, MD, PhD



## When Does Computational Validation Trump Biological Validation?

Many a successful investigator working at the interface between molecular biology, genetics and computation will recognize the imperative to obtain biological validation for computational investigations. Even if they have extensively mined multiple datasets of prior research done by others, experience will have shown that the lack of an additional, novel validation dataset will make it challenging to overcome the reviewers' concerns. This is particularly the case for the top tier, "high impact" journals. Of note, this expectation of additional, novel biological validation will be stated not only by reviewers from a traditional molecular biology or genetics background, but also by many of us in the bio-computational community. I wish to argue here that in many instances, such requirements are the result of an inadequate understanding of the nature of the data being used and their value as compared to a novel incremental dataset. Moreover, such requirements represent a failure of the bio-computation community's confidence in their own methodology and a similar failure in our ability to educate our broader biological investigational community regarding what constitutes a figure of merit in a modern computationally-assisted scientific investigation.

I was recently reminded of this failure when I presented results from work involving our National Center for Biomedical Computing, i2b2 (Informatics for Integrating Biology and the Bedside) during a session at a Keystone meeting on insulin resistance. Using multiple datasets from one of my previous collaborations with the Joslin Diabetes Center in Boston, we had undertaken a simple meta-analysis across multiple experiments conducted by leading investigators in type 2 diabetes involving mouse models or human models of insulin resistance. In collaboration with i2b2 investigator **Peter Park, PhD**, we found

that no gene was significantly expressed across all the murine and human models. Nonetheless we found that in 8 of 17 experiments one gene was differentially expressed—which I thought remarkable given the diversity of heterogeneous mouse and human experiments involved. Yet after my presentation, colleagues expressed skepticism about the validity and interest of these results, given that the analysis brought together so many disparate conditions and organisms. The dominant scientific culture expects novel results to arise only under a highly specific set of conditions in individual investigator's laboratories. However, **Dr. Mitch Lazar** from the University of Pennsylvania came to the podium immediately after my presentation and generously remarked that I had scooped him! His own research (a genome-wide chromatin immuno-precipitation scan) had also revealed the significance of that same gene in insulin resistance and adipogenesis, a result he confirmed in several *in vitro* studies. Dr. Lazar's results will soon be published in a first tier journal—and deservedly so. Yet it would have been very challenging for our purely computational analysis to receive similar treatment.

There are without a doubt several purely computational analyses from which any biological conclusions drawn are suspect. Further experimentation or data are required before any tentative conclusions can be drawn. Equally suspect, however, but far more often published, are biological results from an *in vitro* experiment in a non-human model organism under conditions having little to do with those experienced in the course of human pathology. Nonetheless there is a class of computational investigations that leverage prior, often published data sets, sometimes singly and sometimes together. Can we establish a scientific theory or at least a reliable set of heuristics as to when such investigations are sufficient? Are there conditions when an overwhelming set of "lightly used" previously published data can be re-explored to even greater effect and greater generality and applicability than a narrow set of biological experiments? Are there indeed a set of computational investigations that require no additional biological validation? Those of us who work at the intersection of computation and biology are both the best placed to provide principled answers to these questions and also should be the most motivated to so. Let the games begin. □

### DETAILS

Isaac Kohane, MD, PhD, is the Lawrence J. Henderson Associate Professor of Pediatrics and Health Sciences and Technology at Harvard Medical School; Chair of the Informatics Program at Children's Hospital, Boston; and Principal Investigator for Informatics for Integrating Biology and the Bedside (i2b2) a National Center for Biomedical Computing.