Interview

Isaac Kohane speaks to Sarah Miller, Assistant Commissioning Editor

Isaac Kohane, MD, PhD, is the Lawrence J Henderson Professor of Pediatrics and Health Sciences Technology at Harvard Medical School (MA, USA), where he is also Director of the Countway Library of Medicine, the Children's Hospital Informatics Program and Co-Director of the Harvard Medical School Center for Biomedical Informatics. In addition, he is Director of the Informatics for Integrating Biology to the Bedside (i2b2) center (MA, USA), one of seven NIH-funded National Centers for Biomedical Computing. Dr Kohane is also the Co-Director of the Bioinformatics and Integrative Genomics PhD Training Program at the Harvard-MIT Division of Health Sciences and Technology. He earned his Sc.B in Biology from Brown University (RI, USA), and his MD and PhD in computer science from Boston University (MA, USA). Dr Kohane leads multiple collaborations employing genomics and bioinformatics to study cancer and brain development, and he practices pediatric endocrinology at the Children's Hospital Boston (MA, USA). He is involved in the development of multiple hospital computer systems which can be used as 'living laboratories', linking genetic information with clinical data to study the genetics of disease, whilst respecting patient privacy. Dr Kohane has developed cryptographic health identification systems, automated personal health records and peer-to-peer pathology information networks. He also leads NIH-funded initiatives for translating genomic research into the clinical environment, such the i2b2 initiative.

Isaac S Kohane

Harvard Medical School, 10 Shattuck Street, Boston,
MA 02115, USA
isaac.kohane@tch.harvard.edu

■ Following your MD at Boston & your subsequent move to Children's Hospital Boston, why did you decide to pursue & specialize in informatics & genomic research?

I interrupted medical school to complete a PhD in computer science. Based on the knowledge and data overload that was evident in clinical medicine even in the 1980s, the systematization of knowledge acquisition, data analysis and decision support appeared quite compelling to me.

When microarray technology came of age at the turn of the millennium, it became equally clear that genomic and biological science was going to be limited by our ability to analyze this new torrent of data. With the economies of scale and efficiency in genome-scale measurement driven by the continued advance of biotechnology, it seemed likely this trend was going to continue to increase. The informatics methodological armamentarium – for example, machine learning, ontology development and scaleable data management – was clearly just as applicable to these genome-scale data as to

conventional clinical data. As a result, from around 1999, I became very invested in genomic and bioinformatics research.

■ In your opinion, what have been the most significant developments in the field of pharmacogenomics over the past few years?

In terms of pharmacogenomics, the most significant developments have been in the field of cancer. It became very clear that cancers were ongoing, evolutionary entities, where new mutations were developing or providing resistance to therapy. As a result of much higher-throughput sequencing (i.e., whole-exome or -genome sequencing), we have been able to identify those resistant cells and mutations and thereby change the therapy. I think that this has been a hugely important milestone in the way that we actually think about cancer therapy, and it also makes a real difference today as to how we are treating, and how we are going to be treating, cancers.

Another change which has happened over the last few years is on more of a

promissory note: the emergence of electronic-health record data. In our own i2b2 system or in the Vanderbilt BioVU system, or the eMERGE network that the National Human Genome Research Institute (NHGRI) has funded, we can identify through medical records those patients who have had a variety of adverse or idiosyncratic events, and we can potentially correlate their genomic information with these findings. I say this is a promissory note, because although we have had successes in finding genomic variants of interest, we are yet to have the big discovery which will result in a change in medical practice.

■ **Can you describe your recent work regarding the numerous, incidental genetic abnormalities, or ‘incidentalomas’, uncovered in genomics research?**

It has always intrigued and sort of appalled me that, in medicine, we appear to perform screening studies which appear to come and go as fads. For example, in prostate cancer, prostate-specific antigen screens were previously held up as absolutely essential to the care of middle-aged men. Now, by and large, it has been discredited as a screening study, which means, implicitly, that tens of thousands of men have had their prostates removed unnecessarily, at a great morbidity and cost. Similarly, there is now increasing controversy surrounding mammograms.

So I asked myself: if that’s merely one or two tests, what’s going to happen when we are able to do a million such tests at one time – such as we can today for a few hundred dollars, with an “old-school” SNP microarray, let alone what we can do with whole-genome sequencing? So, I did a very simple back-of-envelope calculation and assumed that genetic tests were far more sensitive and specific than they really were; 100% sensitive and 99.9% specific. If you ran 10,000 such tests in parallel, then 60% of the population would be falsely positive – and that was a conservative estimate. If that were to occur, it would have a terrible impact on pharmacogenomic medicine. Not only would you have patients worried – as they should be – by these false positives, but also doctors would be irritated and worried by all of these consultations with the

worried patients, and overwhelmed by all the additional testing and clinical evaluations that were involved. Then, ultimately, the payers, whether these are governments or insurance companies, would say: “Wait a minute, not only are we not going to pay for genetic tests, but we will not pay for the following procedures/treatments that are being done because of these false-positive genomic tests”.

This realization led me to coin the term the ‘incidentalome’; the biggest ‘ome’ of them all, the ‘-ome’ of all incidental findings. I cautioned that if we do not come to grips with this incidentalome, then we are putting genomic medicine at risk.

One of the key insights is that the reason we have previously been able to do well with much less-sensitive and specific tests, is that we order them on patients in whom there is already a clinical suspicion. It is therefore much more likely than the general population prevalence that they will have this disease, and therefore the test is going to work well. The only way we will be able to apply these tests in the general population is if we first study those general populations to see what factors are confounding these tests in the general population. There are now a lot of well-studied cases in which mutations have occurred in families where they have a strong effect, but in the general population they do not. And so this calls, I think, for large-population studies; and these can, in fact, be enabled by the electronic health records that I previously described.

■ **You are Director of the i2b2 centre; how do you hope this initiative will aid research into pharmacogenomics & genetic disorders?**

Unfortunately, right now, every time a patient comes into the hospital system, the system does not learn anything, except for the individual doctor who saw the patient. i2b2 is supposed to make every healthcare visit an opportunity for the whole healthcare system to learn. The way we did this was to say: “Let’s take all the informational and biological byproducts of each visit, and put those into a de-identified datamart for recurring and ongoing analyses.”

Although using codes and various structured laboratory measurements is very important for obtaining insights, there is a lot more that is written about the patient in the narrative text of the medical record. And so we made a big investment in using natural language processing to collect knowledge for the text. As a result, we were able to identify patients matched to their aforementioned biological samples for genomic studies at approximately a hundredth of the cost of prior studies, and about ten-times faster.

This approach allows us to be nimble, timely and more relevant in the resulting genomic studies. We can, for example, look at populations with asthma and see what SNPs inform us of whether they are resistant to glucocorticoids. Or – and this is a study we completed recently – we can look at patients with rheumatoid arthritis and see if we can find which SNPs predict more serious disease versus a more moderate course of disease.

Another advantage of i2b2 is the multi-layered patient-privacy protection regarding the contents of the electronic health record. This may explain, in part, why i2b2 software, which is free and open-sourced, has been adopted by 60 academic health centres in the USA and 12 internationally, including at least a couple in the UK; I believe the University of Leicester is using it for a coronary cardiac-genomics database.

■ **Do advances such as those described for the i2b2 initiative influence your practice at the Children's Hospital Boston?**

The answer is, as a matter of fact, yes. We have taken a lot of these technologies and applied them to create the Gene Partnership Project at the Children's Hospital, where every single patient's information from their electronic health record gets merged with their genomic data. If there is an important finding, whether it be genetic or pharmacogenomic, we are going to communicate any relevant results in the coming year. What that means, for example, is that if someone is about to get a narcotic for a medical procedure, we will first check if their pharmacogenomic profile predicts whether they are going to be a standard or slow metabolizer of the narcotic.

■ **What challenges do you foresee in the future, both for biomedical computing & the field of pharmacogenomics?**

I suppose that the most important challenge is neither technological nor biotechnological but sociological: how will we obtain consent in a standardized fashion across all patients, so that we, as patients, have control over who gets to look at our data and when? And if we wanted to, how can we be assured that our data will, or will not, be used for research as per our wishes? It is especially important for a heterogeneous sociopolitical landscape, such as the USA or the European Union, to make sure that all of these consent regimes work together so that we can actually conglomerate sufficiently large populations to be able to study rare diseases or very subtle effects of multiple genetic variants.

The second big challenge is that we have to make the knowledge that relates to pharmacogenomic practice freely available, just as textbook knowledge of medicine is. I do have concerns that there are many companies who are trying to make a business out of the interpretation of the pharmacogenomic tests. I worry that this will create impediments for the fullest use of this information.

■ **In your opinion, what new developments would allow the continued implementation of bioinformatics & genomic research into biomedical discovery & improving clinical care?**

Understanding both response to environment and response to medication is going to be essential, so there are two different technologies that are going to have to evolve. Regarding response to the environment, I think we have to look at non-traditional sources of biomedical data, like blogs, Twitter and various apps that allow citizens throughout the world to report on various conditions. We have to figure out how to systematically exploit this. So, on the environmental side we really need to figure out how to crowd source ourselves as consumers. In fact, members of our faculty have already done this for infectious diseases with HealthMap.org.

Also, I think looking at epigenomics and seeing how, for example, the modifications of chromatin are changed in response to drugs – and the environment – will be the source of new biomarkers for response. I think this will be hugely important in pharmacogenomics.

■ **Do you have any concluding remarks?**

Unfortunately we have to ask ourselves a hard question: who is going to pay for all of this? I think in order for governments or insurers to pay for it, we have to redouble our efforts to show the actual clinical value

of pharmacogenomic profiling to make the case for using these technologies.

Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.