Big Ideas, Big Data
HOW GENOMICS WILL CHANGE OUR WORLD

Pierre Meulien, Fiona Brinkman and Jennifer Gardy

A computer-generated image that maps out part of the complex network of interactions between thousands of proteins in cells. Behind each piece of the image is data that unveils secrets of a given protein's role in important cell functions. Red lines denote new interactions more recently uncovered by research. Image courtesy Fiona Brinkman

Genomics is the most dynamic, rapidly progressing field of science of the 21st century. But, already, the societal deliverables of this breathtaking field are moving beyond the wealth of information being gleaned from genomic mapping to the myriad applications of that data for innovation in health care, environment and industry. Every day, researchers in Canada and beyond are leveraging high-performance computing and novel analytical techniques to turn Big Data into Big Ideas across our economy's key sectors.

If the 20th century was that of the computer, the 21st century will be that of biology. In the past decade alone, technological advances have fueled discovery in the field of genomics at a dizzying rate, spurred by the reality that we can now decipher the biological code encrypted in DNA at a speed and cost unthinkable just ten years ago.

But reading genomes—the genetic instruction books of life present in all living things—isn't enough. In order to harness the knowledge we can now glean by reading the genomes of humans, crops, trees and microbes to drive change in our health care, our industries and our environment, that breathtaking wealth of

new information has to be digested and directed. What will drive progress and success in these areas is analysis of these datasets in the context of other variables—markers of human health, geographic and environmental data, biochemical pathways, climate models. In other words, Big Data.

Big Data is an all-encompassing term for any collection of datasets so large and complex that they become difficult to process using traditional data processing applications. Genomics datasets are a prime example, given that a simple bacterium's DNA is millions of letters of code (called base pairs) long, a single human genome contains about three billion base pairs of DNA, and the wheat genome contains a whopping 16 billion base pairs. Since 1982, the number of DNA bases in the world's main gene databank, GenBank, has doubled approximately every 18 months. Combine data about how the genes, proteins and other associated molecules are changing in the organism with environmental or clinical metadata about where the sequences came from and the amount of data coming out of genomics experiments can quickly become hugely challenging to analyze. A simple computational analysis of 22,000 sequences from a microbial genomics experiment involves running 140 trillion DNA sequence comparisons against GenBank.

Despite these challenges, researchers in Canada and beyond are leveraging high-performance computing and novel analytical techniques to harness the power of genetic information and other datasets, turning Big Data into Big Ideas across our economy's key sectors.

Of all the food consumed by Earth's seven billion people, 85 per cent of the calorific value comes from just 12 species of plant; either directly through our intake of cereals, vegetables, and rice, or indirectly through feeding of livestock and fish. These crops have been selected over millennia for their production yield, but we don't know what traits we've unintentionally eliminated from these varieties, from flavor profiles to pest resistance. We have dramatically decreased the biodiversity of our crops, leaving them vulnerable in ways that could substantially impact our food systems.

Cancer strikes two in every five Canadians and when it does, it does so in a very individual manner—no two cancerous growths have exactly the same characteristics, though all are driven by changes in our DNA.

Genomics researchers are tackling this problem from many angles. Some are genetically characterizing older varieties of crops stored in the world's seed banks and reverse engineering biodiversity into modern crops, hoping to enhance their innate resilience. Others are using genomics approaches to understand what happens to crops when they're stricken with fungal infections or pest infestation in order to better protect these precious resources.

Similar approaches are being deployed in the forestry sector, helping Canada to effectively manage the almost 600 million seedlings planted nationwide each year by understanding the combination of genetic and environmental factors that determine the right tree to plant in the right place, and our forests' genetic basis for resilience to pests such as the pine beetle. Canadian researchers are even working toward a genomic catalogue of all our planet's biodiversity, developing a DNA-based identification system capable of reading 'the barcode of life'.

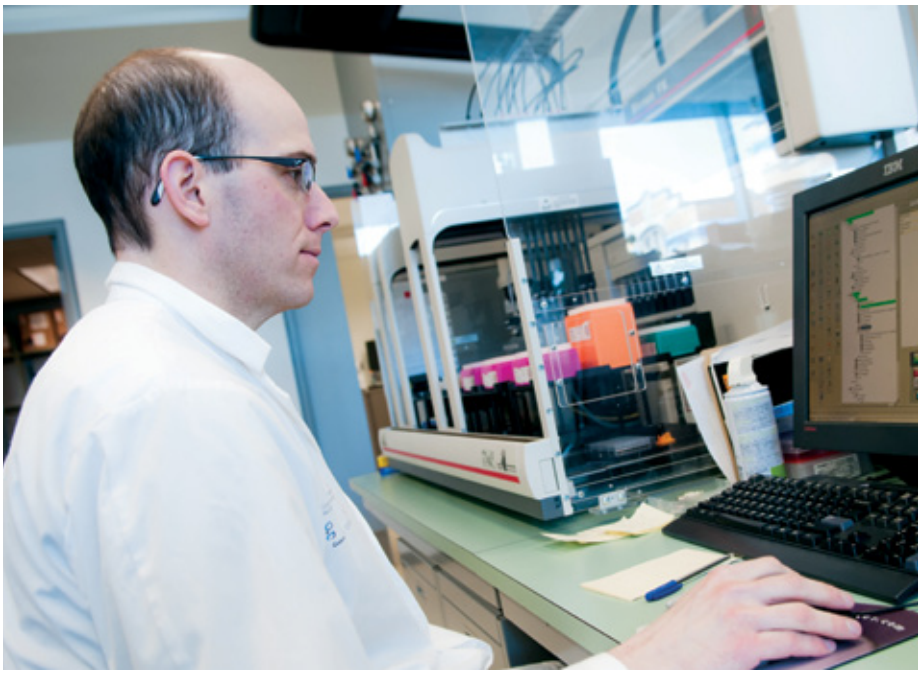
Beyond preserving biodiversity, researchers are also concerned with monitoring the health of our environment—another area in which genomics is transforming the landscape. Many studies have set out to discover "biomarkers"—genetic readouts from samples such as soil and water or from organisms like salmon that signal change, either positive or negative. These analyses require integrating large genomic datasets with environmental measurements across a range of metrics. In Canada, researchers are sequencing the microbial DNA present in clean and polluted water to identify novel biomarkers of water quality that perform better than traditional coliform counts and, if pollution is detected, help track the cause. Genome-wide studies of fish genes, coupled with reproductive data, have been used to better model the impact of steroids on fish populations and forecast future populations, while the recently sequenced Atlantic salmon genome—a Canadian-led project—has yielded data that are informing best practices for aquaculture and breeding.

Earlier and more accurate forecasting of the environmental impacts of toxicants is of great value to environmental risk assessors—improving our ability to know when to act (and when not to). Genomics can also instruct us how to act. Written into the genetic code of many microbes is a remarkable — and highly marketable — ability to digest pollutants, from carcinogenic polychlorinated biphenols (PCBs) to petroleum byproducts. Leveraging these organisms for bioremediation in, for instance, oil sands tailings ponds and oil spills, is a promising new strategy made possible by our understanding of these microbial genomes and the many useful functions they encode.

Cancer strikes two in every five Canadians and when it does, it does so in a very individual manner—no two cancerous growths have exactly the same characteristics, though all are driven by changes in our DNA. Canada is participating in the International Cancer Genomics Consortium, which is sequencing the genomes of over 25,000 tumors of 50 different types to understand how and why they developed. This genomic revolution is already leading to changes in the way cancer is treated. Canadian scientists and clinicians are working hand in hand to sequence tumour DNA from patients, interpreting the resulting genomic data through the lens of biochemical pathways to personalize treatment plans targeting

individual tumours and the genetic changes that caused them. This tailored approach is already showing promising results, and we envision a future in which every patient presenting with a cancer will have their tumour genome sequenced as part of the standard of care.

Genomics isn't just affecting chronic disease. By reading the DNA base pairs in a viral or bacterial pathogen's genome, we are gaining important insights into infectious disease. Canadian researchers have been at the forefront of using new genomics technologies, combined with epidemiological and clinical data, to better understand and control disease outbreaks and epidemics. They can use such data to track outbreaks of foodborne or respiratory illnesses, figuring out how a disease is moving through a community, and quantify how the disease-causing microbes change over time—impacting treatment or vaccination programs. These are all insights critical to designing effective, evidence-based provincial and federal public health initiatives.



A researcher looks at data coming from a next-generation genome sequencing machine at the McGill University-Génome Québec Innovation Centre, one of five such centres supported by Genome Canada across the country. Photo courtesy of Genome Canada

While the opportunities for the Canadian economy presented by genomics and Big Data are clear and multiplying, we still have challenges to overcome. Chief among these is the issue of efficient and effective integration of genomics datasets with other data streams. If done correctly, with appropriate data standards, Canada will be positioned to be leaders in translating Big Data into Big Economic Efficiencies. If not, we risk floundering in a sea of data and missing the potential for significant discovery. Integration requires the collaborative effort of stakeholders across a range of domains, each working together to develop standardized vocabularies and formats for data exchange.

Data exchange and integration must be supported by appropriate data governance to ensure effective, secure and appropriate use of the many datasets. Several initiatives are approaching this issue, including the Global Alliance for Genomics and Health, Europe's ELIXIR initiative, and the international Global Microbial Identifier project. These consortia will develop the policies and practice around data sharing, privacy, and data interoperability, but their recommendations and best practices will still face significant regulatory hurdles on the road to implementation. Interdisciplinary teams of physical and social scientists, lawyers, policymakers and others are convening to address these issues, and government support at levels from municipal to federal will be key to their success.

Data processing analysis will also remain a challenge, especially as our genomic throughput increases and we generate ever-larger datasets. The data from one water quality biomarker study alone would take over 7000 days to analyze on one computer, so a high performance computing infrastructure and network, such as that provided by Compute Canada and CANARIE, is key to enabling discovery in a Big Data world. We also require new tools for data visualization and statistical analysis scaled to the Big Data level and customized for the genomics landscape. This is an area in which Canada, through its strong computational biology research community, can play a lead role in developing the tools of tomorrow.

We are living in the era of biology, with genomics technology driving everything from the market value of a dairy cow, to the drugs a doctor prescribes to treat an individual patient's condition. Genomic data is bringing value and innovation to industries across Canada's bioeconomy, and with the OECD's prediction that the global bioeconomy will reach over a trillion dollars by 2030, our nation, with its unique and significant natural resources footprint, should be primed to claim a significant portion of this. By adopting a Big Data approach to a genomic exploration of the world around us and promoting policies and practices that support data integration, analysis, and appropriate governance, we can translate genomics data into economic efficiencies in a wide range of sectors, positioning Canada as a leader in the new global bioeconomy. P

Fiona Brinkman is Professor, Molecular Biology and Biochemistry at Simon Fraser University. brinkman@sfu.ca

Jennifer Gardy is Senior Scientist at the BC Centre for Disease Control. jennifer.gardy@bccdc.ca

Pierre Meulien is President and CEO of Genome Canada. pmeulien@genomecanada.ca