September 13, 2016

BACKGROUNDER
2015 Bioinformatics and Computational Biology Competition results

Genome Canada, in partnership with the Canadian Institutes of Health Research (CIHR), is pleased to announce 16 projects selected for funding as a result of the 2015 Bioinformatics and Computational Biology Competition. A total of $4 million is being invested in these projects, which will produce next-generation tools and methodologies to deal with the influx of large amounts of data produced by modern genomics technologies, and will provide broad access to these tools to the research community.

BRITISH COLUMBIA

New bioinformatics for new sequencing technologies: Genome characterization and variation detection using long reads

Project leader: Inanc Birol, British Columbia Cancer Agency
Lead Genome Centre: Genome British Columbia
Total funding: $250,000

Gene sequencing has become a routine and cost-effective method for conducting research in many life sciences fields. However, the most dominant sequencing technology today generates millions of short sequences, consisting of 100-300 bases (the building blocks of DNA). These short “reads” have to be assembled in the right order to make sense of the data – the same way a box of jigsaw pieces promises a picture, but only when the pieces are put together in the right order. Longer reads are possible, but they have higher error rates and a large amount of “noise,” leaving researchers to choose between short, high-quality reads with gaps, or long, but noisy, data.

Dr. Inanc Birol of the British Columbia Cancer Agency is a world leader in genome assembly. Now he proposes to develop specialized software to quickly, accurately, and efficiently assemble and analyze long sequence reads. The new tools, which will be available for free online, will allow teams around the world to make faster progress on diverse projects.

Automated analysis of big flow cytometry data

Project leaders: Ryan Brinkman, British Columbia Cancer Agency; Cedric Chauve, Simon Fraser University; Sara Mostafavi, University of British Columbia
Lead Genome Centre: Genome British Columbia
Total funding: $249,994

Flow cytometry (FCM) works by suspending cells in a fluid and then passing that fluid through a laser. FCM-based assays are now widely applied in health research and are important for diagnosing blood cancers, monitoring viral infection (e.g., HIV), vaccine development and stem cell research. As FCM technology has improved over the years the amount of information it delivers per cell has increased five-fold, to the point it is virtually impossible to analyze manually without missing much of the information embedded in those cells. In order to unleash the power of big FCM data, analytical tools and statistical methods must be developed and made widely available.

Drs. Ryan Brinkman (BC Cancer Agency), Cedric Chauve (Simon Fraser University) and Sara Mostafavi (UBC) are developing an easy-to-use, graph-based approach to represent the information generated by FCM as well as new approaches for supervised and unsupervised cell population identification . Among other benefits, users will be able to identify novel patient groups, such as those who don't respond to certain drugs, enabling the delivery of truly personalized medicine.

PathOGIST: Calibrated multi-criterion genomic analysis for public health microbiology

Project leaders: Leonid Chindelevitch, Simon Fraser University; William Hsiao, University of British Columbia; Cedric Chauve, Simon Fraser University
Lead Genome Centre: Genome British Columbia
Total funding: $250,000

Public health agencies protect the population from infectious diseases by monitoring disease-causing agents (pathogens) and intervening in disease outbreaks. Without these concerted efforts, our population's health would be endangered. Today, public health workers can conduct pathogen whole-genome sequencing to investigate risks and determine the source of outbreaks. But the challenge of translating that data into actionable interventions remains.

Drs. Leonid Chindelevitch, Cedric Chauve (Simon Fraser University) and William Hsiao (UBC) are developing a statistically sound and robust computational framework called PathOGIST to enable public health workers and others to quickly classify pathogens into epidemiologically related groups based on sequencing data and generate interpreted genomic reports to inform their actions. PathOGIST will revolutionize how disease outbreaks are managed, ensuring faster responses that will significantly reduce the impact of these outbreaks on both health and the economy.

On Target: Big Data-Informed Software for the Design of cis-Regulatory Regions Controlling Human Gene Expression

Project leader: Wyeth Wasserman, University of British Columbia
Lead Genome Centre: Genome British Columbia
Total funding: $250,000

Some two million Canadians live with a genetic disease. Many of these diseases are not well understood and treatment and care is a challenge; existing treatments are often prohibitively expensive, imposing costs beyond the capacity of the health care system. Gene therapy has the potential for treating numerous genetic diseases. The concept is simple: gene therapy delivers and activates a gene to replace an abnormally functioning gene, thereby restoring the function of that gene. Its application is more complicated, in part because most current therapies do not restrict activation of the delivered gene to affected cells only, potentially resulting in undesired side effects.

Dr. Wyeth Wasserman of the University of British Columbia is leading a team to develop OnTarget, a software application that will allow scientists to design short DNA sequences to specifically activate delivered genes in targeted cells only. Through a user-friendly web-interface, scientists will be able to select and optimize human DNA sequences to target therapeutic gene expression to the right cells for their disease of interest. OnTarget will reduce the costs of designing gene therapy treatments with the potential to increase safety and efficacy. The ultimate beneficiaries will be patients with genetic disease in need of treatment.

Genomic Epidemiology Application Ontology (GenEpiO)

Project leaders: William Hsiao, University of British Columbia; Andrew G. McArthur, McMaster University; Fiona S.L. Brinkman, Simon Fraser University
Lead Genome Centre: Genome British Columbia
Co-Lead Genome Centre: Ontario Genomics
Total funding: $250,000

Infectious disease outbreaks have significant impacts on human health, agri-food production, animal health and the economy. Ineffective public health responses can result in outbreaks that spread diseases like the Zika virus and food-borne illnesses, with enormous impacts on health and high economic costs. DNA sequencing provides the complete “fingerprint” of a microbe, enabling an unprecedented tracing of how infectious diseases spread. When outbreaks become global, however (think SARS, or microbes resistant to antimicrobials) data needs to be shared across public health organizations securely and efficiently. Unfortunately, data is often held in institution-specific formats, making it difficult, time consuming and costly to do so.

Drs. William Hsiao (UBC), Andrew G. McArthur (McMaster University) and Fiona Brinkman (Simon Fraser University) will improve data integration and sharing of infectious disease and

antimicrobial resistance information across public health agencies, with the Genomic Epidemiology Application Ontology (GenOpiO). The platform will enable public health workers to share outbreak-related information faster and to perform more powerful analyses, helping to reduce the negative health and economic impact of disease outbreaks.

ONTARIO

Enhanced and Automated Visualization of Complex Data

Project leader: Paul C. Boutros, Ontario Institute for Cancer Research
Lead Genome Centre: Ontario Genomics
Total funding: $250,000

Modern genomics research generates massive amounts of data. But these data sets are too big and complex to be useful on their own. Researchers must first analyze and interpret biological data to better understand them and turn them into meaningful information. This information can then be used to help solve real-world problems, such as developing new tools or strategies to better diagnose and treat patients, increasing crop yields or monitoring the environment. Increasingly, the ability of the human end-user to interpret the data is the key factor limiting researchers from delivering these much-needed solutions more quickly.

Dr. Paul C. Boutros of the Ontario Institute for Cancer Research is leading a team developing ways of making "big data" results more easily understood by improving the way it is visualized and interpreted. The team will create interactive visualization tools that will integrate tightly with databases scientists already use routinely. The team will use crowdsourcing to capture the best visualization ideas from a broad community of scientists, graphic designers and citizen-scientists. The project will build on the human brain's ability to interpret images, to make the conclusions of biological data more readily accessible and accelerate the rate of biological discovery and innovation.

Consolidated epigenetic landscape for congenital, developmental and childhood disorders

Project leaders: Michael Brudno, Rosanna Weksberg, Hospital for Sick Children
Lead Genome Centre: Ontario Genomics
Total funding: $249,900

Epigenetics is the study of both genetic and external factors, such as environmental exposure or lifestyle choices by parents or grandparents, which affect gene expression. Epigenetic disruptions play a key role in disease. Finding epigenetic biomarkers, however, is complicated by the complexity of epigenetic signaling in cells or tissues, as well as the fact that many different genetic disorders, such as pediatric developmental disorders, can show similar clinical symptoms. Despite the wealth of data being generated by new technologies, there is a dearth of diagnostic tools that can consolidate epigenetic data collected by diverse groups using

different experimental platforms. These tools are essential to relate molecular patterns to clinical presentation.

Drs. Michael Brudno and Rosanna Weksberg of Toronto's Hospital for Sick Children are developing a novel web-based resource for analyzing epigenetic datasets together with complete clinical information, focusing on developmental disorders such as intellectual disability and autism. Their system will provide a rich context for exploring epigenetic dysregulation in a growing number of childhood epi-genetic diseases.

Dockstore: A platform for sharing cloud-agnostic tools with the research community

Project leader: Vincent Ferretti, Lincoln Stein, Ontario Institute for Cancer Research
Lead Genome Centre: Ontario Genomics
Total funding: $250,000

An unintended consequence of the development of genomics has been the proliferation of massive datasets, making analysis increasingly difficult. A further problem is the lack of standardization in how analysis tools are packaged, described and executed across computer environments. Drs. Vincent Ferretti and Lincoln Stein of the Ontario Institute for Cancer Research, in collaboration with Dr. Brian O'Connor of the University of California, Santa Cruz, have developed a web application called the Dockstore, which addresses the challenge of encapsulating and sharing bioinformatics tools so that they can be moved from environment to environment.

Now the researchers are adding key features to the Dockstore to continue to enhance and evolve the platform. They will also integrate bioinformatics tools and workflows from the Global Alliance for Genomics and Health (GA4GH) for redistribution to the larger research community and will work with collaborators to facilitate the registration of their high-quality tools into the Dockstore. Finally, the researchers will work with other projects to enable sharing of tools across genomic repositories. These activities will drive increased usage of the Dockstore, thereby increasing tool sharing among scientists in fields as diverse as agriculture, energy and human health.

Kamphir: a versatile framework to fit models to phylogenetic tree shapes

Project leader: Art F.Y. Poon, Western University
Lead Genome Centre: Ontario Genomics
Total funding: $205,365

Phylodynamics is a new and rapidly growing field that combines epidemiology and computational biology to combat infectious disease outbreaks. The field stems from the concept of phylogeny, in which a tree represents how different populations – of virus infections, for example – are related through a series of common ancestors. The genetic

similarities among populations are used to reconstruct these ancestral relationships back in time. This is particularly important for viruses, which evolve so quickly that each infection becomes genetically unique within weeks or months of being transmitted from the previous host. Consequently, the virus phylogeny can be used to estimate how the infections spread through the host population. Phylodynamics has already had an enormous impact on our understanding of outbreaks including HIV, hepatitis C virus, and Ebolavirus. Further progress is stymied, however, by simple models that can't accommodate large data sets.

Dr. Art F.Y. Poon of Western University, Ontario, is developing a completely new approach to phylodynamics that adapts a method from pattern recognition to enable computers to "see" the shared features of different tree shapes. This approach will have an unprecedented capacity for more realistic models and larger data sets, improving global public health initiatives for infectious disease management and eradication.

ePlants Pipeline and Navigator for Accessing and Integrating Multi-Level 'Omics Data for 15 Agronomically Important Species for Hypothesis Generation

Project leader: Nicholas Provart, University of Toronto
Lead Genome Centre: Ontario Genomics
Total funding: $250,000

In the past five years alone, huge amounts of data have been generated for 15 plant species important for Canada, including poplar, maize, rice, barley, wheat, soybeans and tomatoes. Being able to efficiently use these data will be key to improving and managing these crops to feed, shelter and power a world of 9 billion people by the year 2050.

The ePlant Framework, developed under a previous Genome Canada grant, permits researchers to easily see where and when a gene is "active" and whether there are natural genetic variants that might allow it to do its "job" better; populated only with one species, it now needs data from more species. Lead researcher Dr. Nicholas Provart (University of Toronto) plans to develop an ePlant Pipeline to facilitate the ability to create any ePlant, based on genomic or exome sequence data. The ePlant Navigator will permit cross-cultivar and cross-species comparisons, supporting robust hypothesis generation. Easy access to these data sets will enable researchers to explore genetic diversity, gene expression, and other data for important genes towards crop improvement.

Rapid, accessible genome assembly using long read sequencing

Project leader: Jared Simpson, University of Toronto and Ontario Institute for Cancer Research
Lead Genome Centre: Ontario Genomics
Total funding: $250,000

DNA sequencing technology has progressed from sequencing single reference genomes at great cost and time, to the current era of inexpensive, high-throughput short read sequencing. The

emerging “third generation” of DNA sequencing technology offers the prospect of putting long read genome sequencing in the hands of more researchers and enabling new applications, through portable instruments that will decentralize sequencing technology.

Dr. Jared Simpson of the University of Toronto is developing robust and efficient genome assembly software that is easy to use, to match the capabilities of these emerging sequencing instruments. The software will target biologists and other end users of sequencing who don’t have substantial bioinformatics expertise.

QUÉBEC

Computation of cell-specific microRNA:mRNA regulatory networks enables the design of efficient RNAi-based therapeutics

Project leaders: François Major, Université de Montréal, Thomas Duchaine, McGill University
Lead Genome Centre: Génome Québec
Total funding: $250,000

Just as your home has a security system to protect against invaders, so has your body. This bodily security system is a process of shutting down genes known as gene silencing, and can fight both external invaders, such as viruses, and internal dysfunctions, such as those that lead to cancer.

Gene silencing is based on microRNAs, a class of RNA. Recent developments in biochemistry techniques allow researchers to create artificial microRNAs and introduce them in cells. Drs. François Major (Université de Montréal) and Thomas Duchaine (McGill University) are developing computational tools to identify the genes involved in specific diseases and design treatments based on microRNAs that can silence a set of predetermined genes simultaneously. Once developed, the software will be available to the scientific community through easy-to-use web-interfaces and applications.

A toolkit for genome-wide association studies in bacteria

Project leaders: Jesse Shapiro, Université de Montréal; Luis Barreiro, CHU Sainte Justine
Lead Genome Centre: Génome Québec
Total funding: $250,000

There is a pressing need for better diagnosis of antibiotic-resistant infections. Diagnostics allow doctors and veterinarians to prescribe the right antibiotics in a timely fashion. In many cases, this is essential for minimizing morbidity and mortality from the infection and reducing the costs of antibiotic-resistant infections, currently estimated at $500 million/year. One of the most significant barriers is the length of time it takes to culture bacteria.

Molecular diagnostics offer the potential for faster, point-of-care tests, but require knowing which genes or mutations to look for, a challenge when the genes involved in antibiotic resistance are still unknown. Drs. Jesse Shapiro (Université de Montréal) and Luis Barreiro (CHU Sainte-Justine) are developing software to identify the bacterial genes or mutations associated with antibiotic resistance – or any other bacterial trait of interest – using genome-wide association studies. This software will provide the missing link between bacterial genome sequencing and reliable, rapid molecular diagnostics. Researchers, clinicians and veterinarians will be able to make informed decisions about treatment, reducing costly and unnecessary treatments and improving human and animal health.

Computational methods and databases to identify small RNA-binding molecules regulating gene expression

Project leaders: Jérôme Waldispühl, Nicolas Moitessier, McGill University
Lead Genome Centre: Génome Québec
Total funding: $249,999

Messenger RNA (mRNA) are RNA molecules that carry messages from DNA as part of gene expression. Discovered only in 2002, riboswitches appear to be an ancient and widespread class of mRNA elements. One particular riboswitch, the flavine mononucleotide (FMN)-activated riboswitch, is a promising target for antibiotics. This riboswitch is critical to the production of riboflavin (vitamin B_2) in bacteria but not in humans. Activating the bacterial FMN riboswitch would shut down this biosynthetic pathway, preventing the growth of bacteria without affecting human health. Very likely, several other riboswitches are potential targets for novel small molecule drugs. These riboswitches and small molecules binding to them are yet to be discovered.

Searching for small molecules can be tedious and time-consuming. Virtual high throughput screening is both quicker and less costly than any experimental approaches. Profs. Jérôme Waldispühl and Nicolas Moitessier, both at McGill University, are developing the computational infrastructure and technology needed for genome-wide screening of riboswitch elements and the identification of novel small molecules that activate them.

Crowdsourcing Genomic Databases

Project leaders: Jérôme Waldispühl, McGill University; Olivier Tremblay-Savard, University of Manitoba
Lead Genome Centre: Génome Québec
Co-Lead Genome Centre: Genome Prairie
Total funding: $250,000

The online availability of large genomic datasets, annotations, and broad knowledge in biology is an essential source of information that has made a huge contribution to the rapid expansion of this field. To remain relevant for today's research, these resources need to be maintained

and constantly updated with new biological information. Despite their fundamental role in the discovery pipeline, however, support for key genomic databases is in jeopardy.

Drs. Jérôme Waldispühl (McGill University) and Olivier Tremblay-Savard (University of Manitoba) are taking genomic databases to social media, by developing a web and mobile crowdsourcing platform for curating, maintaining and updating genomic databases. The system will be tailored to high school and university students and instructors. The completion of this project will demonstrate the broad impact of crowdsourcing on solving critical technical and financial challenges faced for maintaining key biological research resources.

ATLANTIC

Rapid prediction of antimicrobial resistance from metagenomic samples: data, models, and methods

Project leaders: Robert Beiko, Dalhousie University; Andrew G. McArthur, McMaster University
Lead Genome Centre: Genome Atlantic
Co-Lead Genome Centre: Ontario Genomics
Total funding: $250,000

Antimicrobials (antibiotics), have been central to combating infectious disease for nearly a century. However, their effectiveness is slipping due to the increase in antimicrobial resistance (AMR). There is an increasingly urgent need to know more about AMR to better understand its consequences and monitor its presence in the environment, agri-foods industry, individual patients, and on a population level. Being able to analyze the genomes of resistant microorganisms is essential, but slow and costly to do one at a time. Metagenomics allows genetic profiling of microbes as a community, but datasets are huge and contain much irrelevant data. Currently, there is no software designed to specifically predict AMR profiles directly from metagenomic data, which would enable more rapid AMR profiling and aid prioritization of candidate genes for further research.

Drs. Robert Beiko of Dalhousie University, Andrew G. McArthur of McMaster University, and Fiona Brinkman of Simon Fraser University are leading a project to develop new software and database tools that will provide a near-instantaneous picture of AMR organisms in a sample, aiding AMR research and responding to AMR threats impacting both agri-food production and public health.