GenomeCanada

February 4, 2019

BACKGROUNDER

2017 Bioinformatics and Computational Biology (B/CB) Competition Results (Multi-sector stream projects)

Genome Canada is pleased to announce 25 projects selected for funding as a result of the 2017 Bioinformatics and Computational Biology Competition (B/CB). A total of $23.1 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.

Bioinformatics and Computational Biology is specifically recognized as a shared priority among Canada's research funders, as it will be the key to extracting meaning from increasingly complex large data sets.

ALBERTA

A comprehensive analytical toolkit and high-performance genome browser for rapid, reliable and in-depth characterization of bacterial genomes
Project leaders: Paul Stothard, University of Alberta; Gary Van Domselaar, Public Health Agency of Canada-National Microbiology Laboratory
Genome Centre: Genome Alberta
Total funding: $940,000

Sequencing DNA has become easier and less expensive, leading to rapidly expanding genome databases and a better understanding of life in general. Yet while a genome sequence contains important clues about the history and characteristics of an organism, the data generated by current sequencing technologies does not emerge in an easily interpreted form. Instead, results are often produced as small fragments, akin to the words or phrases that constitute a more meaningful but jumbled recipe or description. Moreover, the definitions of many of the words and phrases found in DNA sequence information are often not clear. Genomes, like recipes, are easier to understand if they are written in a familiar language, laid out in logical order, and supported with helpful images.

Many scientists today have difficulty processing their sequence data into a form suitable for their research applications and look to bioinformaticians for help. However, this reliance on others is inefficient, particularly in the field of microbial genomics for which the pool of skilled bioinformaticians is small.

Drs. Paul Stothard of the University of Alberta and Gary Van Domselaar of the Public Health Agency of Canada are overseeing the development of *Proksee*, a software system that will allow researchers to convert raw bacterial sequence data into high-quality, richly described, and easily interpreted whole-genome assemblies (i.e., microbial life recipes). These powerful analytical resources will be freely available to all researchers, including those working in small-sized Canadian companies without the expertise or resources to transform raw sequencing data into a meaningful form.

By helping Canada's scientific community translate DNA sequence information into new insights into the biology of important bacterial species, the sophisticated and user-friendly *Proksee* platform will support broad-ranging scientific discoveries and innovations, including the development of safer and healthier foods, new techniques for dealing with pollution, and new methods of manufacturing valuable products.

ATLANTIC

Antimicrobial Resistance: Emergence, Transmission, and Ecology (ARETE)
Project leaders: Robert Beiko, Dalhousie University; Fiona Brinkman, Simon Fraser University
Genome Centres: Genome Atlantic, Genome British Columbia
Total funding: $1.4 million

Bacteria are becoming increasingly resistant to antimicrobial agents, causing concern for the agri-food industry, where overuse and mismanagement of antibiotics have played a significant role. There is no single solution that will reverse this trend, nor reinvigorate the antimicrobial discovery pipeline. A better understanding of the genes that make bacteria resistant and how they spread is, therefore, a priority worldwide. A key element in the transmission of resistance is the sharing of resistance genes among pathogenic bacteria. Resistant bacteria also move between habitats, such as agricultural soil and farm animals, but we need a better understanding of key transmission points in order to prioritize surveillance and regulation activities.

Drs. Robert Beiko of Dalhousie University and Fiona Brinkman of Simon Fraser University are leading a large team made up of academic and government partners who are seeking to determine which genes are being shared, which bacteria are sharing genes, and how bacteria are moving between habitats. They are developing bioinformatics algorithms and software that will shift how we look at antimicrobial resistance (AMR) from a static "snapshot" to a dynamic view of AMR transmission. They will validate their tools using thousands of genomes of Salmonella, E. coli, and other pathogenic bacteria collected by partners in the Public Health

Agency of Canada and Agriculture and Agri-Food Canada. The integrated software pipeline that will result from this project will be open-source and freely available.

The rigorous bioinformatics framework developed through this project will enable more-informed approaches to minimize the risks posed by resistance and foster a national framework to apply genomics and bioinformatics to the "farm to fork" continuum for AMR**.**

BRITISH COLUMBIA

AnnoVis: Annotation and visualization of *de novo* genome and transcriptome assemblies
Project leader: Inanc Birol, British Columbia Cancer Agency
Genome Centre: Genome BC
Total funding: $1 million

Today, it is possible to know more about a species than ever before by decrypting its genome (DNA) and transcriptome (RNA.). However, processing this data into a coherent form and interpreting it is more difficult than simply collecting it. Characterizing the genes of a particular species plays a critical role in interpretation. For instance, the presence of certain genes in a given genome may be indicative of the wood quality of local forestry products or provide insights into the infectious patterns of a pathogen in the food chain. However, comprehensive and accurate gene identification and annotation depends on the completeness and correctness of the underlying information.

The utility of genomic and transcriptomic data is determined by two factors: experimental design and the analysis methods used. The former is well covered; Dr. Inanc Birol of the BC Cancer Agency is focusing on the latter goal, which remains a challenge. He is building methods to improve the quality of assembled genomes and transcriptomes by detecting misassembled sequences. He will also develop tools and resources to facilitate gene annotation and predict their function. Further, he and his team will also develop visualization tools to assess the quality of assemblies and their associated annotations.

The tools will be made available through Dr. Birol's software portal and will help researchers world-wide better understand the world around us at the genomic and transcriptomic levels.

Proteogenomics-Improved and –Guided Quantification Pipeline (PIGQpipe): Targeted Proteomics with Internal Proteogeno-typic Peptide Standards to Quantify Variants Identified by Proteogenomic Experiments
Project leaders: Christoph Borchers, Yassene Mohammed, University of Victoria
Genome Centre: Genome BC
Total funding: $556,472

Measuring proteins in biofluids such as urine or blood is critical for diagnosing, treating and monitoring disease in humans. Currently researchers and clinicians worldwide use mass spectrometry (MS) arrays to conduct multiple and parallel reaction monitoring (MRM/PRM) for

this task. MRM/PRM approaches enable precise, accurate quantitation of proteins in samples, but still depend on reference human protein-sequence databases, which may be missing disease-specific genomic variants. To increase use of these approaches, therefore, the pace of assay development and validation has to increase, as does the pace of the actual sample analysis. As well, data analysis has to be standardized and automated to enable robust, large-scale implementation.

Right now, there is no sophisticated and intuitive software framework to help researchers design high-throughput mass spectrometry-based assays, process the data and interpret the results, making the process of using mass spectrometry difficult, laborious and time consuming.

Drs. Christoph Borchers and Yassene Mohammed of the University of Victoria are developing a new software pipeline, PIGQpipe, that will integrate the researchers' own experiment-derived data with data from public online databases to generate a "one-stop shop" for all aspects of mass-spectrometry assay design, data analysis and interpretation. Using a single web-based interface, PIGQpipe will connect, integrate and automate assay workflow (including selecting targets and optimal experimental conditions), using data to calculate protein concentration values, statistically evaluating differences between treatment groups or disease conditions and presenting the data.

By increasing the productivity, scale, quality and scope of research in this area, PIGQpipe should enable new research discoveries, including tools for precision medicine; facilitate and expedite the adoption of mass-spectrometry technologies into the clinic; and build expertise in the scientific community.

Machine learning methods to predict drug resistance in pathogenic bacteria
Project leaders: Leonid Chindelevitch, Maxwell Libbrecht, Simon Fraser University, and Jesse Shapiro, University of Montreal
Genome Centre: Genome BC
Total funding: $1,000,000

Treatment options for people infected with antibiotic-resistant bacteria today are limited, and may become even more so over time. The risk of a post-antibiotic era, in which even minor infections or routine medical procedures could be fatal, is real. Beyond the personal costs, the economic burden of drug resistance is high, estimated to cost more than $1 billion each year in North America alone.

Genomic screening of pathogens to determine their identity and responsiveness to antibiotics could be both more accurate and efficient than other ways of diagnosing infectious disease and choosing the appropriate treatment. The challenge, though, lies in two areas: analyzing the large amount of data produced by this sequencing to understand the genomic factors underlying drug resistance, and creating an accurate model to predict drug resistance based on sequencing data.

Drs. Leonid Chindelevitch and Maxwell Libbrecht of Simon Fraser University and Jesse Shapiro of University of Montreal are developing computational tools based on machine learning, which will be able to unravel the complex relationships between bacterial genome sequences and antibiotic resistance, based on bacterial genome data currently available in public databases as well as from their collaborators.

The project, which builds on two projects funded in the 2015 bioinformatics and computational biology competition, will result in two tools: a comprehensive predictive model for resistance to a variety of drugs for specific bacteria, which will update itself as new data become available; and a user-friendly web interface to enable researchers, including clinical and public health researchers, to securely upload and analyze genomic data from pathogenic bacteria and accurately predict drug resistance. The tool looks to boost research in drug resistance and development of better clinical tools to the ultimate benefit of infectious disease patients.

Illuminating the dark matter of the metabolome with convolutional neural networks
Project leaders: Leonard Foster, University of British Columbia; David Wishart, University of Alberta
Genome Centres: Genome BC, Genome Alberta
Total funding: $500,000

The human body contains hundreds of thousands of naturally occurring chemicals, and is exposed to hundreds of thousands more during daily life. These chemicals, called metabolites, are involved in key biological processes in humans and other large organisms. Metabolomics is the scientific discipline that allows scientists to measure these small molecules at a large scale. Metabolomics can reveal molecules involved in disease, produce diagnostic and prognostic tests and predict how patients will respond to specific prescription drugs. Much has been accomplished with current computational tools, which can only identify a small fraction of the metabolites in a sample. Better computational tools that could identify the remaining metabolites would dramatically accelerate the pace of metabolomics research.

Drs. Leonard Foster of UBC and David Wishart of the University of Alberta are developing computational tools based on an artificial intelligence technique known as “deep learning” to handle the huge amount of data generated by metabolomics experiments. The first tool, DeepMet, will increase the number of molecules that can be identified in metabolomics experiments. The second, MetUnknown, will help assign chemical structures to molecules that are as of yet unknown. Together, these tools will help shine a light on the majority of the metabolome that is overlooked by current tools.

DeepMet and MetUnknown will be available in three different formats so that scientists in many different areas can use the tools without any specialized training. Ultimately, these tools could help researchers identify therapeutic targets for complex diseases such as cancer and develop new tests to help physicians personalize medical treatments.

Global scale metabolic pathway reconstruction from environmental genomes

Project leader: Steven Hallam, University of British Columbia
Genome Centre: Genome BC
Total funding: $1.2 million

We may think it's all about us humans, but for more than 3.5 billion years, microorganisms have been the dominant form of life on earth. Over the past decade, high-throughput sequencing and mass spectrometry have generated information about the DNA, RNA, proteins and metabolites found in these microbes. Together, this provides information about their function and identity, linking them to a wider range of ecosystem functions at the individual, population and community levels. There is, however, a paucity of scalable software tools to mine, monitor and interact with environmental datasets. This is particularly frustrating at a time when we as a society are grappling with global climate change, as microbial communities offer a virtual blueprint to rebuild our global future in more sustainable ways.

Dr. Steven Hallam of the University of British Columbia is leading the development of the Environmental Genome Encyclopedia (EngCyc), a compendium of microbial community metabolic blueprints supported by high-performance software tools. Access to EngCyc will be through a web portal and support user-defined blueprint construction in an automated and scalable manner, thus enabling gene and pathway discovery. Its data exploration options will power knowledge creation and translation.

The combination of the web portal and software tools will enable the research community, both nationally and internationally, to more effectively explore and harness the hidden powers of microbial communities, to the benefit of the biorefining, mining and energy sectors, among others.

Bioinformatics Tools to Enable Federated, Real Time Genomic Epidemiology Data Sharing and Analysis in a One Health Framework
Project leaders: William Hsiao, University of British Columbia; Gary Van Domselaar, Public Health Agency of Canada
Genome Centres: Genome BC, Genome Prairie
Total funding: $1,164,488

Infectious diseases such as influenza, Ebola, Listeriosis and Salmonellosis, as well as other pathogens, can devastate both animal and human lives, damage economies and paralyze trade. The One Health approach recognizes that the health of humans, animals and the environment are closely intertwined, requiring a collective approach to effectively detect, respond to and prevent outbreaks.

Genomics has transformed the detection and characterization of pathogens, expediting the development of diagnostic tests and vaccines and expanding our knowledge of pathogens. Utilizing genomics data for real-time surveillance among partners, though, is difficult, because of inconsistent contextual information associated with genomic samples, lack of a trusted and

secure data-sharing platform, and inadequate tools for localized and collaborative genomic analyses.

Drs. William Hsiao of UBC and BC Centre for Disease Control Public Health Laboratory, and Gary Van Domselaar of the Public Health Agency of Canada are creating new data-sharing platforms and data-processing tools to enable real-time, multijurisdictional data sharing, allowing researchers – and their software systems – across sectors to communicate information faster, more accurately and more securely. The tools will be tested by three Canadian collaborating centers and with international partners to ensure their fitness and usability.

The software tools developed in this collaboration aim to transform how infectious disease data is shared and analyzed, leading to better monitoring of the emergence and spread of pathogens in wildlife, food and food animals, and the environment, which will reduce disease burden, prevent agri-food trade embargoes and minimize costly food product recalls. It will improve communication between public health and agricultural institutions, and enhance collaboration at the provincial, national and international levels. Ultimately, this project should improve the health and well-being of Canadians.

Development and implementation of bioinformatics tools for HIV and HCV phylogenetic monitoring platforms
Project leaders: Jeffrey B. Joy, Julio S.G. Montaner, University of British Columbia
Genome Centre: Genome BC
Total funding: $1.2 million

The World Health Organization (WHO) has set targets to eliminate HIV/AIDS as a pandemic by 2020 and Hepatitis C virus (HCV) by 2030. In the developed world, where epidemics are well managed, it is becoming increasingly difficult to identify remaining pockets of ongoing HIV and HCV transmission, making it difficult to reach and go beyond these targets. To facilitate reaching these targets we need new bioinformatics tools that can identify communities at high risk of continuing HIV and HCV infection and be used to support the delivery of public health interventions and healthcare services.

Phylogenetic monitoring – harnessing large, rapidly growing sequence databases – offers a solution for identifying clusters of HIV or HCV diagnoses in near real-time. Developing and implementing such a system also requires new bioinformatic tools to facilitate sharing the results of analyses with public health officials while maintaining patient confidentiality and privacy. Such tools would also need to be able to take advantage of next-generation sequencing data.

Drs. Jeffrey B. Joy and Julio S.G. Montaner at the University of British Columbia are developing and implementing such a system. It will serve as a national, near-real-time phylogenetic monitoring platform capable of incorporating based on next-generation sequencing data, coupled with bioinformatic tools to securely present and distribute results to public health agencies. They are developing new methods of monitoring epidemics and infections. These

tools will be freely available to the research community and allow more effective targeting of public health interventions, moving us more rapidly toward reaching WHO targets to control and eliminate HIV and HCV.

ONTARIO

BridGE-SGA: A novel computational platform to discover genetic interactions underlying human disease

Project leaders: Charles Boone, University of Toronto; Chad L. Myers, University of Minnesota
Genome Centre: Ontario Genomics
Total funding: $990,910

The ability to sequence the entire human genome at increasingly lower cost has led to a fundamental change in biomedical research. But there is a gap between the amount of data available and our ability to understand and interpret that data. Addressing this gap is essential to realize the promise of precision medicine.

Dr. Charles Boone and Dr. Brenda Andrews of the Donnelly Centre for Cellular and Biomolecular Research at the University of Toronto, and Dr. Chad Myers of the University of Minnesota, have worked together to discover that a significant part of our inability to interpret genomic data likely stems from the reality that disease generally arises from complex genetic interactions. While all humans essentially have the same set of genes, most have around five million unique genetic variants. The effect of any one variant depends on its interactions with other variants. So we need to understand not just the millions of genetic differences that affect gene function, but also how all those genes interact with each other. Current computational methods and technologies lack the statistical power to do so.

Drs. Boone, Andrews, Myers have developed the first complete genetic interaction map for any organism, and have built a computational method, BridGE, to discover genetic interactions. The team is now working to develop an innovative computational platform for genome sequencing data, BridGE-SGA, to enable the discovery of disease-associated genetic interactions from large-scale human genotype data. Their goal is to discover genetic interactions for a variety of diseases. Identifying and understanding these key genetic interactions will improve our ability to interpret data from whole genome sequencing and identify novel gene targets for drug discovery and development.

Computational tools for Data-Independent Acquisition (DIA) for quantitative proteomics and metabolomics

Project leaders: Anne-Claude Gingras, Lunenfeld-Tanenbaum Research Institute; Hannes Röst, Donnelly Centre for Cellular & Biomolecular Research, University of Toronto
Genome Centre: Ontario Genomics
Total funding: $1,000,000

When cells lose control over their own behaviour or communication with other cells, diseases such as diabetes or cancer can arise. Protein and small molecule metabolites are responsible for cells' behaviour, so identifying and quantifying these molecules is key to understanding how disease happens and how to prevent it.

Mass spectrometry has become the workhorse for proteomics and metabolomics. Drs. Anne-Claude Gingras of the Lunenfeld-Tanenbaum Research Institute and Hannes Röst of the Donnelly Centre for Cellular & Biomolecular Research at the University of Toronto are working with a technology called Data-Independent Acquisition (DIA), in which the mass spectrometer systematically identifies and quantifies the proteins and metabolites present in a sample. DIA has been shown to improve quantitative accuracy, reproducibility and throughput over other methods. Since its introduction, however, this approach has only been applied to small-scale studies and in a relatively small number of laboratories. Limitations to this method are due to the lack of user-friendly software that could enable a scalable analysis of the complex data generated in large-scale biomedical and medical research.

The project builds on the team's proven strength in DIA data analysis and software development and will result in an integrated set of tools available under an open-source license. To encourage uptake of these tool, documentation, webinars and workshops will be made available to potential users. The results of the project could have long-lasting impact on the health sector in Canada by facilitating research into the root causes of disease and assisting with clinical questions such as patient stratification.

SYNERGx: a computational framework for drug combination synergy prediction
Project leader: Benjamin Haibe-Kains, Princess Margaret Cancer Centre
Genome Centre: Ontario Genomics
Total funding: $1,032,702

When just one drug is used to treat cancer, the patient may not respond, or may develop resistance to it. Combination therapy, where two or more drugs are used in treatment, is more likely to be successful. Yet, it is impossible to test all drug combinations in clinical trials due to the high cost of required resources and certain ethical considerations. Computational techniques are therefore required to model the large amount of available data to improve current cancer treatment strategies and propose more efficient combinations of drugs.

Dr. Benjamin Haibe-Kains of the Princess Margaret Cancer Centre is developing SYNERGx, a new computational platform that will integrate multiple pharmacogenomic datasets. These datasets will be used to predict possible combinations of known drugs that can act in synergy, meaning that their combined therapeutic efficacy is greater than the sum of their individual effects.

The platform will implement analytic tools to improve modeling of synergistic drug effects. Users will have access to highly curated drug-combination pharmacogenetics data and an open-source machine-learning pipeline for drug synergy prediction. SYNERGx will also implement a

new way to optimize drug-screening studies to identify novel synergistic combinations that can be further validated in preclinical studies and then in clinical trials.

SYNERGx will provide an efficient way to leverage massive investments in pharmacogenomics studies by allowing the integration of otherwise disparate datasets. It represents a major step forward in the design of new therapeutic strategies for cancer.

Software for Peptide Identification and Quantification from Large Mass Spectrometry Data using Data Independent Acquisition
Project leaders: Bin Ma, University of Waterloo; Michael Moran, Hospital for Sick Children
Genome Centre: Ontario Genomics
Total funding: $925,987

Precision medicine gives patients the opportunity to tailor medical and treatment decisions at the individual level to maximize outcomes and minimize adverse effects. It can be used to treat a wide variety of diseases, including cancer. Decisions are often based on the presence and quantity of biomarkers such as proteins in the blood or tissue samples.

Advances in mass spectrometry instruments have made it feasible to discover and measure protein biomarkers, but researchers lack the necessary bioinformatics software to analyze the data. Drs. Bin Ma of the University of Waterloo and Michael Moran of the Hospital for Sick Children are developing this software to enable more sensitive and accurate protein identification and quantification from the mass spectrometry data generated using a method called data independent acquisition (DIA). They expect that their software will significantly increase the total number of proteins identified and quantified in comparison to existing DIA analytical software. It will be especially effective with post-translational modifications (PTMs), which are critical biomarkers in a proteins' function and degradation.

The free availability of the software to academic labs coupled with its superior performance can help health researchers discover and trace disease biomarkers. Within the next decade, the software could become an indispensable tool for many proteomics labs performing DIA analysis throughout the world. The new software may also help commercial partners create value-added new products, services and jobs.

Ultimately, this will lead to improvements in human health and reduction in healthcare costs by enabling early disease detection and diagnosis and by facilitating the selection of optimal treatment for individual patients.

CReSCENT: CanceR Single Cell ExpressioN Toolkit
Project leaders: Trevor Pugh, Princess Margaret Cancer Centre; Michael Brudno, The Hospital for Sick Children
Genome Centre: Ontario Genomics
Total funding: $1,000,000

Tumours are complex mixtures of cancer, immune, and normal cells that interact and change during treatment. The interplay of all three types of cells can dictate development of cancer over time, as well as response or resistance to treatments. Recent advances in microfluidic and DNA sequencing technologies have enabled researchers to simultaneously analyze tens of thousands of single cells from complex tissues, including tumours. Interpreting these data is challenging, due to the lack of high-quality reference sets of each cell type in the body and a lack of methods to link these data back to tumour biology.

Drs. Trevor Pugh of the Princess Margaret Cancer Centre and Michael Brudno of The Hospital for Sick Children are developing the CanceR Single Cell ExpressioN Toolkit (CReSCENT), a scalable and standardized set of novel algorithmic methods, tools, and a data portal deployed on cloud computing infrastructure. To allow comparison of cells in cancerous and healthy tissues, the system will aggregate single-cell genomic data generated by cancer researchers and connect them to international reference data generated by experts from around the world as part of the Human Cell Atlas. This data sharing and aggregation system is a key differentiating factor in CReSCENT that will increase researcher productivity by accelerating execution and comparison of computational methods, as well as providing contextual data for understanding how cells behave within tumour tissues.

This platform, which will be useable by any researcher on any computing platform, will assemble a crucial data resource to navigate the upcoming wave of single cell cancer genomics research. CReSCENT will bring together researchers across a broad spectrum of scientific areas and disease types and increase the impact of data generated across research programs. In the long term, this system will pave the way for novel single cell diagnostics and discovery of new drug strategies for improved health care.

Dockstore 2.0: Enhancing a community platform for sharing cloud-agnostic research tools
Project leaders: Lincoln Stein, Ontario Institute for Cancer Research; Mark Fiume, DNAstack
Genome Centre: Ontario Genomics
Stream: 1 (health)
Total funding: $875,269

With Genome Canada support, Dr. Lincoln Stein of the Ontario Institute for Cancer Research successfully developed Dockstore, a system that enables complex computational biology algorithms to be run reliably and reproducibly across multiple platforms. It has been adopted as the leading packaging technology by the Global Alliance for Genomics and Health and is now used by numerous third-party bioinformatics groups. Dockstore was acquired by Canadian company DNAstack and Dr. Stein is now working with DNAstack's Dr. Mark Fiume to maximize the utility of Dockstore.

The aim of these enhancements is to promote greater collaboration and sharing among computational biology software developers. Specifically, the enhancements will make Dockstore easier to use, make its packages more powerful and expressive, increase its interoperability and enable these packages to run more easily on a wide range of systems and

hardware architectures. The bioinformatics and computational biology community will benefit from this software, while the research results derived from it that are reproducible, portable and reusable.

From ePlants to eEcosystems: New Frameworks and Tools for Sharing, Accessing, Exploring and Integrating 'Omic Data from Plants
Project leaders: Nicholas Provart, University of Toronto; Jörg Bohlmann, University of British Columbia
Genome Centres: Ontario Genomics, Genome BC
Total funding: $1,000,000

Major advances in plant biology over the past decade are in large part thanks to new technologies for DNA sequencing and phenotyping (i.e. mapping the physical expression of genetic traits). The resulting datasets allow researchers to determine how different plants develop and respond to changes in their environment. Yet, in order to take advantage of the tremendous amount of new data, innovative tools are required to integrate and visualize the number of individual data points in different datasets. The original ePlant system, developed as part of a previous Genome Canada effort, integrates many data types but was not configured for phenotype data. Amongst its many applications, phenotype data provide important information on traits of interest to plant breeders.

Drs. Nicholas Provart of the University of Toronto and Jörg Bohlmann of the University of British Columbia are developing a new module to integrate the wide variety of data available, including ecosystem data, phenotypes and genotypes into ePlant. This will be done for the already existing ePlant species and any new ePlant species to be developed as part of this project. The researchers will also open the ePlant system to the research community to build a larger ePlant ecosystem of information. This online system will act as a resource where plant biologists will be able to share their datasets.

Ultimately, these tools can help to accelerate the task of identifying useful genes to feed, shelter and power a world of nine billion people by the year 2050.

Extracting Signal from Noise: Big Biodiversity Analysis from High-Throughput Sequence Data
Project leaders: Sarah Adamowicz, Paul Hebert, University of Guelph
Genome Centre: Ontario Genomics
Total funding: $507,231

Surveying biodiversity is critical for environmental health and for managing natural resources. It helps to assess the impact of resource development, but also to identify pests, invasive species, and pathogens in a rapid and cost-effective manner. It is essential to Canada's economic growth in the forestry, agriculture, and fishery sectors and to decision-making in public health. Genetic methods of surveying biodiversity, such as high-throughput sequencing, are being broadly adopted, but bioinformatics has not kept pace with the data being generated. In addition,

current methods are geared toward bacteria and similar organisms, rather than multi-celled plants and animals that need monitoring as well.

Drs. Sarah Adamowicz and Paul Hebert, along with colleagues from the University of Guelph, are creating new bioinformatics tools that will facilitate the rapid and accurate processing of DNA data resulting from high-throughput sequencing. The tools will enable the simultaneous analysis of bulk samples, which are made up of many different species. It will include a de-noising tool to detect errors; a method to cluster DNA sequences into species-like units to permit biodiversity analysis; and a method for assigning sequencing data to higher taxonomic categories to unlock functional biological information. The team will combine these various tools into a biodiversity informatics pipeline that can be incorporated into existing web-based platforms for uptake by a broad variety of users.

The new biodiversity informatics tools will support large-scale biodiversity research by academics; efficient, accurate, and cost-effective environmental assessments for the mining and pulp-and-paper industries; enhanced capacity and accuracy of regulation; and more rapid and accurate biodiversity data for government and private-sector decision-makers.

QUEBEC

Bioinformatics tools for integrative 3D epigenomics
Project leaders: Mathieu Blanchette, Jacek Majewski, Jérome Waldispühl, McGill University
Genome Centre: Génome Québec
Total funding: $1,122,405

Since the sequencing of the human genome at the beginning of the millennium, scientists have made great strides in understanding the role genes play in our identity, differences among individuals and our susceptibility to diseases such as cancer. Bioinformatics and computational tools are critical in visualizing and analyzing genetic data, but they fall short when it comes to three-dimensional analysis of how DNA is folded onto itself to fit into the nucleus, known as its chromatin architecture. This 3D structure defines the genome's normal function during early cell differentiation and development and is known to drive many developmental disabilities as well as cancer.

There is a pressing need for a new generation of computational tools that integrates our knowledge of these 3D states, to help researchers make optimal use of the rapidly increasing amounts of data produced by modern DNA sequencing. Dr. Mathieu Blanchette and his team at McGill University will create and release improved computational and statistical tools for analyzing 3D data in their native 3D context. Their new tools will be integrated with the team's 3D visualization platform that will help scientists explore the data, build new hypotheses and test them in rigorous statistical frameworks.

The results of this project will be made widely available through open-source software that will enable statistically robust analysis of individual and groups of 3D genomic data; provide a

virtual reality-based 3D genome browser supporting integrated visualization of genomic data; and include a toolset for integrative mining of genomic and epigenomic data in their 3D genome context. A lay version of the visualization platform will also be used for community outreach through exhibits in schools and museums.

Epigenomics Secure Data Sharing Platform for Integrative Analyses (EpiShare)
Project leaders: Guillaume Bourque, Yann Joly, McGill University
Genome Centre: Génome Québec
Total Funding: $1,000,000

Epigenetics is the study of reversible modifications on the genetic material of cells, affecting gene expression mechanisms. They are partly inherited, and partly imputable to environment and life habits. Thanks to advances in next-generation DNA sequencing and multiple international efforts, a vast amount of human epigenetic data is now available to researchers. This data can provide explanations of, and insights into, the interpretation of genome-wide association studies (GWAS), the study of genetic variants in an individual genome. Accessing this data, however, can present a significant challenge. This is due in part to very large file sizes stored across multiple locations, and the fact that data needs to be stored behind secure control mechanisms, to help protect research participants' privacy.

The Global Alliance for Genomics and Health (GA4GH) has already developed tools and standards for sharing genomic data securely and ethically. Extending on this framework, Prof. Guillaume Bourque and Prof. Yann Joly of McGill University will develop mechanisms to make the process of accessing and analyzing epigenomic data more flexible while addressing the ethical, security and privacy aspects of data sharing. To accomplish this, they are developing the EpiShare platform, so that researchers can avoid having to download and clean up data and metadata from different epigenomic projects, a time-consuming effort that makes inefficient use of computing resources. Instead, a web portal will make data more easily discoverable and enable the launch of multi-omics analyses on these controlled-access datasets at their storage location. The EpiShare platform will also ensure that, in the process, participants' privacy is maintained.

TooT Suite: **Predication and classification of membrane transport proteins**
Project leader: Gregory Butler, Concordia University
Genome Centre: Génome Québec
Total funding: $600,000

Increasing population + improved standard of living = a threat to the adequacy of our food supply. Indeed, by 2050, when the world population is expected to reach nine billion people, agricultural production will need to increase 60-70% to feed all these people. But with most of the world's arable land already in production, the solution has to involve improving the yields from both crops and animals.

Plants and animals are part of a complex ecosystem, co-habiting symbiotically with microbial communities, called microbiomes, that live in, on or near them, affecting their health and growth and, therefore, their productivity as a food source. Scientists currently use genomics and metagenomics to study these microbiomes, to better understand how microbiome-host interactions affect that health and growth. These interactions happen by exchanging chemical compounds, facilitated by transport proteins, which carry the compounds across the membranes of a cell.

Dr. Gregory Butler of Concordia University is developing *TooT Suite,* a way to annotate the membrane transport proteins both in an organism, be it plant or animal, and in a microbiome, thus providing information about potential interactions between them. *TooT Suite* will be an open-source set of easy-to-use bioinformatics tools that will help genomics researchers in agriculture to better understand these interactions. It will work with different existing classification systems by predicting the protein's most appropriate term for each of these systems, thus overcoming a lack of consistency. *TooT Suite* will open an era of agricultural research based on system-level thinking and metagenomics that is needed to address future food supply and food security challenges.

Bioinformatics and Artificial Intelligence to leverage predictive models of dairy production
Project leaders: Abdoulaye Baniré Diallo (Université du Québec à Montréal), Marc-André Sirard (Université Laval)
Genome Centre: Génome Québec
Total funding: $1,004,258

Dairy is big business in Canada, with some 1.4 million cows, most of them in Québec and Ontario, responsible for $6.17 billion in net revenues each year. Dairy in Canada is also unique in that milk production is subject to a quota system and supported by more than 13,000 farms, mostly family-run. Optimizing productivity, maximizing resources and limiting expenses are directly associated with farm profitability. Innovations such as genetic selection, increased management efforts and control over production variables have helped maximize profitability during the past several years. Production data such as milk yield, milk components, animal characteristics and management conditions are available for the past 40 years but have never been integrated with genomic data.

Abdoulaye Baniré Diallo of the Université du Québec à Montréal and Marc-André Sirard of Université Laval are developing tools that will perform this integration, driving the development of new management practices and allowing a precise lifetime productivity estimate for individual cows. Their data mining and machine learning toolkits will deliver predictive models of dairy production that will influence management practices and optimize dairy farm profitability. The project takes advantage of the advances in genomics that allow an increasingly detailed genetic profile to be established for each individual dairy cow, at relatively low cost, optimizing decisions about feeding, reproduction, therapeutic treatments and caregiving, and will be applicable to farms across Canada. It will also be a stepping stone to including new selection traits such as response to various housing conditions, heifer growth and adaptation to

robotized milking systems, to breeding programs. This research is performed with a multisectorial and transdisciplinary research team composed of scientists from Université du Québec à Montréal, Université Laval and McGill University, in partnership with Valacta and My Intelligent Machines (Mims).

Precision Medicine in Cellular Epigenomics
Project leaders: Celia Greenwood, Lady Davis Institute for Medical Research; Karim Oualkacha, Université du Québec à Montréal
Genome Centre: Génome Québec
Total funding: $660,512

Much of the risk for disease lies in our genes, but other factors also contribute. Sometimes, environment or exposures modify how genes operate. One way that gene activity is altered is by methylation of the DNA, a process whereby a methyl molecule attaches itself to a cytosine nucleotide in DNA. Specific methylation patterns are necessary for normal development and altered methylation plays a role in human diseases such as cancer and autoimmune diseases. It has recently become technically and financially feasible to measure DNA methylation at single-nucleotide resolution on a large scale across the genome, using a method called bisulfite sequencing. But the data from this sequencing often have many missing or imprecise measures, making them difficult to interpret and limiting the potential of such studies to describe the role of epigenetics in disease.

Drs. Celia Greenwood of the Lady Davis Institute for Medical Research and Karim Oualkacha of the Université du Québec à Montréal have assembled a team of experts to develop an algorithm and software package to analyze large-scale, high-dimensional DNA methylation data, so that we can profit from the hidden potential of these data. The initial focus of the methods and package will be scleroderma, a debilitating autoimmune disease leading to scarring across multiple tissues with unpredictable treatment response.

The team has support from a patient-advocacy organization, Scleroderma Quebec and from two Montreal companies, one that develops machine-learning methods and one that provides a high-performance computational platform. Understanding of DNA methylation and its contribution to disease will be revolutionized through these methods and software package.

Next-generation molecular docking leveraging artificial intelligence techniques to understand large-scale ligand binding data sets
Project leader: Rafael Najmanovich, Université de Montréal
Genome Centre: Génome Québec
Total funding: $500,000

Nothing in nature exists in isolation, from the smallest molecule to the largest trees or animals. Everything interacts. Dr. Rafael Najmanovich of the Université de Montréal is concerned with the smallest molecule end of things – specifically the interactions between small molecules and the proteins that govern most cellular metabolism and signaling. He wants to understand the

"molecular recognition" that happens between these small molecules and proteins to better understand the biological processes in order to develop new drugs.

Ligands (in general small-molecules) bind to proteins within surface cavities. The prediction of these interactions is done through docking simulations. Docking methods are widely used and accepted today but have yet to reach the level of realism, speed and accuracy required fulfill their full potential in drug design. Dr. Najmanovich's laboratory has already developed FlexAID (Flexible Artificial Intelligence Docking), which outperforms widely used docking methods. Now, they are developing next-generation docking software that will increase the biological accuracy with which researchers can model docking processes, speed up simulations to rank small molecules and focus on specific protein families that are important in drug design.

The docking methods being developed will create open-source software that will help understand, from a structural point of view, drug discovery, and enhance our understanding of molecular recognition. They will also help advance pharmaceutical research to bring about social and economic benefits for Canadians.

An integrative platform for metabolomics and systems biology
Project leaders: Jianguo Xia, Guillaume Bourque, McGill University; Pierre-Étienne Jacques, Université de Sherbrooke
Genome Centre: Génome Québec
Total funding: $1,094,607

Although a relative newcomer to the 'omics family, metabolomics – the comprehensive study of small molecules or metabolites in biofluids such as blood or urine – is being increasingly applied together with other 'omics technologies to understand complex diseases and biological processes. To deal with the resulting big data challenges, easy-to-use and high-performance bioinformatics tools are urgently required for raw data processing, interpretation, and integration with other 'omics data.

Drs. Jianguo Xia and Guillaume Bourque of McGill University, together with Dr. Pierre-Étienne Jacques of the Université de Sherbrooke, are building on their past success with existing web-based platforms for human health 'omics analysis to develop SystemsAnalyst, a new-generation computational platform integrating resources from Compute Canada and public cloud with the latest visual analytics technologies. Their goal is to develop a powerful one-stop shop to enable efficient, transparent, and reproducible analysis of large amount of data to support systems metabolomics and multi-omics data integration.

SystemsAnalyst will directly benefit clinicians and laboratory-based biologists by providing a powerful, user-friendly platform to address current bioinformatics gaps in metabolomics and multi-omics studies. Case studies on malaria and inflammatory bowel disease will facilitate the identification of important biomarkers, key pathways and biological processes that play important roles in these diseases, ultimately assisting in the development of new treatment strategies. The project will train at least 10 highly qualified personnel who are urgently needed

in multi-omics data analysis and modeling. SystemsAnalyst can be easily applied to studying other complex diseases as well.

Development and Validation of a Web-Based Platform for Environmental Omics and Toxicology
Project leaders: Jianguo Xia, Niladri Basu, McGill University
Genome Centre: Génome Québec
Total funding: $1,047,507

Environmental risk assessment is rapidly moving to 'omics tools and systems biology approaches to evaluate the impact of stressors on animal growth, reproduction and survival. Despite great interest among stakeholder groups, it is clear that those in the environmental life sciences cannot adequately deal with the type and amounts of data that are rapidly emerging. There are no accepted and standardized bioinformatics tools to organize, analyze, visualize and interpret 'omics data from key study species.

Over the past decade, Dr. Jianguo Xia of McGill University has developed a series of web/cloud-based tools for comprehensive 'omics data analysis, visualization and interpretation. These powerful and user-friendly tools have proven highly popular within human health community with thousands of users around the world. A similar omics tool suite is urgently needed to support the growing numbers of omics studies in environmental life sciences, including in sectors such as mining, agriculture, forestry, aquaculture and water quality management.

Now Drs. Xia and her colleague Dr. Basu are developing eco.OmicsAnalyst, an intuitive, cloud-based tool to support such data analysis and visualization, beginning with 12 key ecological indicator species covering fish, birds, mammals and invertebrates. To refine and validate the tools, they will embark on a series of case studies within a particularly important class of stressors, chemical pollution. Their work will be driven by key end-users from across government and academia. A technical guidance document will be produced to facilitate end-user uptake.

Due to the research team's close relationships with stakeholders, there is a high likelihood of uptake of the new tool and of overcoming existing barriers for handling 'omics data across the environmental life sciences. Among the benefits will be improved data quality, more efficient decision making and improved resource utilization.