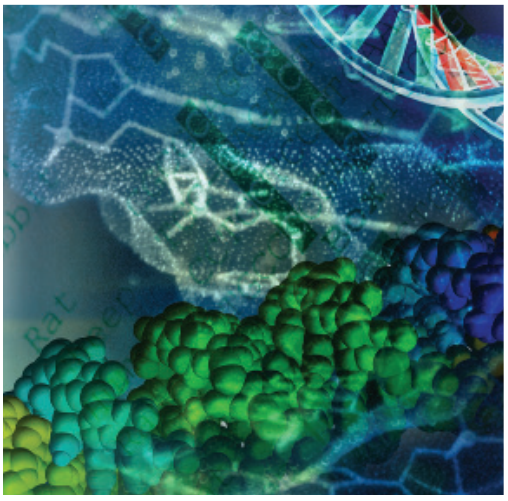
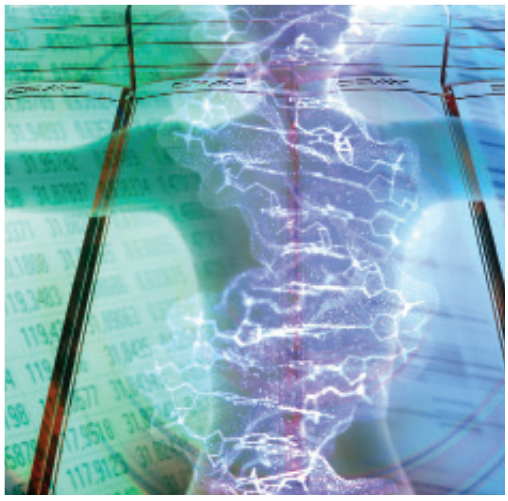
CCBGM BIANNUAL MEETING SEPTEMBER 10-11, 2018

The Mayo Civic Center
30 Civic Center Dr SE | Rochester, MN 55904





CENTER PROPRIETARY

NSF

ILLINOIS

MAYO CLINIC

CENTER FOR COMPUTATIONAL BIOTECHNOLOGY AND GENOMIC MEDICINE (CCBGM)

THE CENTER, MISSION, AND VISION

Center for Computational Biotechnology and Genomic Medicine (CCBGM)

The Center for Computational Biotechnology and Genomic Medicine is a National Science Foundation Industry/University Cooperative Research Center (I/UCRC) that brings two universities with unique resources in computational and biological sciences—the University of Illinois at Urbana-Champaign and Mayo Clinic—together with industry to work on joint projects that will address real-life problems of interest to both industry and academia. The Center offers:

- An approach to the big data problem that spans all of its key elements, from analytics to actionable intelligence, in a broad, comprehensive manner.
- Biological expertise spanning everything from human genomics to crop and animal sciences, along with expertise in computing systems and algorithms (from HPC to the cloud and special-purpose acceleration).
- A strong track record of working with industry in the multidisciplinary domains of computing, biotechnology, and health science.

Center Mission

- To contribute to the nation's research infrastructure base by developing long-term partnerships among industry, academia, and government.
- To leverage NSF funds with industry to support graduate students performing industrially relevant research.

Center Vision

- To expand the innovation capacity of our nation's competitive workforce through partnerships between industry and universities.

AGENDA

Monday, September 10th

8:00–8:30am	**Continental Breakfast**
8:30-9:15am	**IAB Closed-Door Meeting**
9:15-9:30am	**Opening Remarks**
10:30–11:40am	**State of the CCBGM Report** | **RAVI IYER**, Ph.D., Center Director, Univ. of Illinois **LIEWEI WANG**, Ph.D., Center Co-Director, Mayo Clinic
9:45-10:30am	**NSF I/UCRC Presentation** | **DEE HOFFMAN**, Ph.D., CCBGM Evaluator
10:30-10:50am	**MORNING BREAK**
10:50-12:00pm	**Four Funded Project Updates/Quad Chart Report** *(10 minute highlights with 5 min LIFE Forms)* **1** | **From Analytics to Cognition: Taking Genomic and Other Multi-omic Science to the Bedside** | **RAVI IYER,** Ph.D., Professor, Electrical and Computer Engineering, Univ. of Illinois, **RICHARD WEINSHILBOUM**, M.D., Professor, Medicine and Pharmacology, Mayo Clinic **2** | **Information-Compression Algorithms for Genomic Data Storage and Transfer** **OLGICA MILENKOVIC,** Ph.D., Associate Professor, Electrical and Computer Engineering, Univ. of Illinois **3** | **Improving the Accuracy of Genomic Variant Calling Through Deep Learning** **DEMING CHEN,** Ph.D., Associate Professor, Electrical and Computer Engineering, Univ. of Illinois **4** | **Scaling the Computation of Epistatic Interactions in GWAS Data** **LIUDMILA MAINZER,** Ph.D., National Center for Supercomputing Applications (NCSA), Univ. of Illinois
12:00-1:00pm	**LUNCH**
1:00-2:00pm	**Proposals for IAB Consideration** *(30 minutes – 20 min. presentation, 5 min. Q&A, 5 min. LIFE Forms)*
2:00-2:30pm	**Invited Presentation – Imaging Panel** | Moderator Dr. Kiaran McGee
2:30-2:50pm	**BREAK**
2:50-4:15pm	**Project Pre-Proposals** *(15 min. presenation, 5 min. Q&A)*
4:10-4:30pm	**Technical Roadmap Discussion**

AGENDA

4:30-4:35pm **Poster Session Preview** | Students briefly introduce themselves and their posters

4:35-5:00pm **IAB Meeting** | Administrative Issues/Discussions (if needed)

5:00-5:10pm **"Genome: Unlocking Life's Code" – A Smithsonian Institution Exhibit – what to expect** | JAYMI WILSON

5:15-6:15pm **Visit the Genome Exhibit at the Rochester Art Center** (in the Mayo Civic Center)

6:15-7:15pm **Poster Session/reception – all funded projects and proposed projects – LIFE FORMS**

Tuesday, September 11th

7:30-8:00am **Arrival and Continental Breakfast**

8:00-9:00am **LIFE Form Review and Discussion** | All Participants
DEE HOFFMAN, Ph.D., CCBGM Evaluator, Moderator

9:00-9:30am **Live Response and Q&A for Projects**

9:30-11:00am **IAB – Closed-Door Meeting**, IAB Members and NSF
Ask the Directors – Q&A for new industry partners

11:00-11:30am **IAB Report Out, Discussion**

11:30-11:50am **Action Items and Plans for Next Meeting**

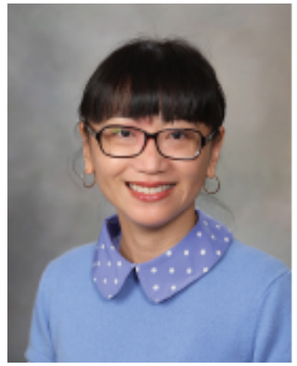
11:50-12:00pm **Summary and Closing Remarks**
RAVI IYER, Ph.D., Center Director, Univ. of Illinois
LIEWEI WANG, M.D., Ph.D., Center Co-Director, Mayo Clinic

12:00pm **Adjourn with Box Lunches**

CENTER LEADERSHIP

Ravishankar K. Iyer, Center Director,
University of Illinois at Urbana-Champaign

Professor Iyer is the George and Ann Fisher Distinguished Professor of Engineering at the University of Illinois at Urbana-Champaign. He holds joint appointments in the Department of Electrical and Computer Engineering, the Coordinated Science Laboratory (CSL), and the Department of Computer Science, and serves as Chief Scientist of the Information Trust Institute. Iyer has led several large successful projects funded by NASA, DARPA, and NSF, as well as industry. He currently co-directs the CompGen Center at Illinois. Funded by NSF and partnering with industry leaders, hospitals, and research labs, CompGen is building a new computational platform to address both accuracy and performance issues for a range of genomics applications. Professor Iyer is a Fellow of the American Association for the Advancement of Science, the IEEE, and the ACM. He has received several awards, including the AIAA (American Institute for Aeronautics and Astronautics) Information Systems Award, the IEEE Emanuel R. Piore Award, and the 2011 Outstanding Contributions Award from the Association for Computing Machinery's Special Interest Group on Security for his fundamental and far-reaching contributions in secure and dependable computing. Professor Iyer is also the recipient of the degree of Doctor Honoris Causa from Toulouse Sabatier University in France.



Liewei Wang, Center Co-Director, Mayo Clinic

Dr. Wang is a Professor in the Department of Molecular Pharmacology and Experimental Therapeutics (MPET) at Mayo Clinic Rochester. She is the Co-PI of the Mayo-NIH Pharmacogenomics Research Network (PGRNP), a network focused on identifying and understanding the molecular mechanisms underlying variation in drug response. Her research involves the application of high-throughput omics technologies to identify biomarkers that might help us individualize cancer therapies, understand mechanisms of drug action, and facilitate target identification and drug development. Dr. Wang is also a co-leader of the Mayo Pharmacogenomics Program within the Mayo Center for Individualized Medicine, where many of the endeavors involving personalized medicine have occurred. She is the recipient of a Mayo Clinic Alumni Association Edward C. Kendall Award and the Astellas Award in Translational Pharmacology from the American Society of Pharmacology and Experimental Therapeutics. She is also deeply involved in graduate education and is the Associate Director for the Mayo MPET graduate program.

CENTER FACULTY AND RESEARCHERS

See the Center's website at **ccbgm.illinois.edu/people** for information about the research areas of these faculty and researchers.

University of Illinois at Urbana-Champaign

Deming Chen, Professor in Electrical and Computer Engineering, Research Professor in Coordinated Science Laboratory, Affiliate Professor in Computer Science

Mohammed El-Kebir, Assistant Professor, Department of Computer Science

Nigel Goldenfeld, University Swanlund Chair in Physics, Director of NASA Astrobiology Institute for Universal Biology, Leader of Biocomplexity Theme at the Carl R. Woese Institute for Genomic Biology

Mikel Hernaez, Director Computational Genomics, Institute for Genomic Biology

Matthew Hudson, Professor of Bioinformatics in the Department of Crop Sciences, Director of the High Performance Computing for Biology Group (HPCBio), Associate Chief Scientist at NCSA, and affiliate of the Carl R. Woese Institute for Genomic Biology and the Department of Entomology

Wen-mei W. Hwu, Professor and Sanders-AMD Endowed Chair in Electrical & Computer Engineering, Chief Scientist of the Parallel Computing Institute, PI of NSF Blue Waters Petascale project

Eric Jakobsson, Professor Emeritus, Molecular & Cellular Biology; Director, National Center for Biomimetic Nanoconductors

Angela Kent, Professor in Natural Resources and Environmental Sciences, College of Agricultural, Consumer and Environmental Sciences (ACES)

Zhi-Pei Liang, Franklin W. Woeltge Professor of Electrical and Computer Engineering, Co-chair of Integrative Imaging theme at Beckman Institute for Advanced Science and Technology

Alexander E. Lipka, College of Agricultural, Consumer and Environmental Sciences, Assistant Professor of Biometry in the Department of Crop Sciences

Steven Lumetta, Associate Professor of Electrical and Computer Engineering and Computer Science

Liudmila Sergeevna Mainzer, Technical Program Manager at NCSA Genomics, Research Assistant Professor at the Carl R. Woese Institute for Genomic Biology

Ruby Mendenhall, Associate Professor in Sociology and African American Studies, Affiliate of the Carl R. Woese Institute for Genomic Biology and Institute for Computing in Humanities, Arts, and Social Sciences

Olgica Milenkovic, Professor in Electrical and Computer Engineering

Klara Nahrstedt, Director of Coordinated Science Laboratory, Ralph and Catherine Fisher Professor in Computer Science

Idoia Ochoa-Alvarez, Assistant Professor of Electrical and Computer Engineering

Gene E. Robinson, University Swanlund Chair, Director of the Carl R. Woese Institute for Genomic Biology, Director of Bee Research Facility

Sandra Rodriguez-Zas, Professor of Animal Sciences, Affiliate of the Carl R. Woese Institute for Genomic Biology

Saurabh Sinha, Professor in Computer Science, Affiliate Faculty of the Carl R. Woese Institute for Genomic Biology, Biophysics, and Entomology

Monica Uddin, Associate Professor of Psychology, Faculty Affiliate of the Carl R. Woese Institute for Genomic Biology

Shobha Vasudevan, Associate Professor of Electrical and Computer Engineering

Tandy Warnow, Founder Professor of Computer Science and Bioengineering, Affiliate Faculty of the Carl R. Woese Institute for Genomic Biology

Bryan White, Professor in Animal Sciences and the Carl R. Woese Institute for Genomic Biology, Director of Mayo Clinic/Illinois Strategic Alliance for Technology-Based Healthcare

Derek Wildman, Professor in Molecular and Integrative Physiology, Leader of Computational Genomic Medicine research theme at the Carl R. Woese Institute for Genomic Biology

Mayo Clinic

Mariet Allen, Assistant Professor of Neuroscience, Senior Research Fellow of Mayo Clinic Florida

William Bobo, Professor of Psychiatry, Chair, Department of Psychiatry, Mayo Clinic Florida

Nicholas Chia, Assistant Professor with appointments in Surgery, Health Sciences Research, and Biomedical Engineering, Associate Director for Microbiome Program in Center for Individualized Medicine (CIM)

CENTER FACULTY AND RESEARCHERS

Travis Drucker, Information Technology Specialist, Team Lead in Bioinformatic Systems, Center for Individualized Medicine

Nilufer Ertekin-Taner, Associate Professor of Neurology and Neuroscience at Mayo Clinic Florida.

Mark Frye, Professor of Psychiatry, Chair of the Department of Psychiatry and Psychology, Director of the Mayo Clinic Depression Center

Matthew Goetz, Professor of Oncology, Professor of Pharmacology, Chair of Mayo Clinic Breast Cancer Disease Oriented Research Group, Deputy Director of Mayo Clinic Breast SPORE

Asif Hossain, Manager of Information Technology Service Delivery, Genomics and Bioinformatics Services

James Ingle, Professor of Oncology, College of Medicine

Clifford R. Jack Jr., Professor of Radiology, Center for Advanced Imaging Research, Alzheimer's Disease Research Center

Krishna Rani Kalari, Associate Professor of Biomedical Informatics, Lead Computational Biologist in the Center for Individualized Medicine and Pharmacogenomics Research programs

Suraj Kapa, Associate Professor of Medicine, Cardiovascular Medicine

Richard Kennedy, Associate Professor of Medicine, Associate Consultant II-Research, Division of General Internal Medicine, Department of Internal Medicine

Eric Klee, Assistant Professor of Biomedical Informatics, Lead Bioinformatician for Clinomics Program in Center for Individualized Medicine, Co-Director and Head of Bioinformatics for Genome Sequencing Lab

Jean-Pierre Kocher, Chair of Division of Biomedical Statistics and Informatics, Director of Bioinformatics Program, Vice Chair of Department of Health Sciences Research at Mayo Clinic in Arizona, Associate Chair of Department of Biomedical Informatics at Arizona State University

Konstantinos Lazaridis, Professor of Medicine, Mayo Clinic College of Medicine; Consultant, Division of Gastroenterology and Hepatology, Department of Medicine; Everett J. and Jane M. Hauck Associate Director, Center for Individualized Medicine

Kiaran McGee, Professor of Medical Physics, Assistant Professor of Biomedical Engineering, College of Medicine; Director of the Imaging Biomarker Program, Center for Individualized Medicine

Heidi Nelson, Chair of the Department of Surgery, Microbiome Program Director in Center for Individualized Medicine

Inna Ovsyannikova, Professor of Medicine

John Port, Professor of Radiology, Associate Professor of Psychiatry, College of Medicine; Consultant, Neuroradiology

Carlos Sosa, Senior Analyst Programmer, Advanced Analytics Services

George Vasmatzis, Assistant Professor of Laboratory Medicine at Mayo Medical School, Consultant in Molecular Medicine, member of Mayo Clinic Cancer Center, Co-Director of Biomarker Discovery Program in Center for Individualized Medicine

Prashanthi Vemuri, Associate Consultant II, Radiology Research; Associate Professor of Radiology, College of Medicine

Liewei Wang, Professor of the Department of Molecular Pharmacology and Experimental Therapeutics, Associate Director of the Pharmacogenomics Translational Program in Center for Individualized Medicine

Richard Weinshilboum, Professor of Molecular Pharmacology and Experimental Therapeutics and Medicine, Director of Pharmacogenomics Program in Center For Individualized Medicine

Eric Wieben, Director of Medical Genome Facility in Center for Individualized Medicine, Director of Office of External Research Collaborations

Mathieu Wiepert, Section Head, Information Technology, Laboratory and Extramural Applications, Instructor in Biomedical Informatics, College of Medicine

Gregory Worrell, Professor of Neurology; Chair, Division of Clinical Neurophysiology; Associate Chair, Neurology Research, Department of Neurology

Curtis Younkin, Analyst/Programmer, Department of Information Technology

Below is a succinct description of the funded projects of the Center. On subsequent pages are the full descriptions of the projects. Also listed below are descriptions of pre-proposals and proposals, which will be presented for discussion at this meeting.

Currently Funded Projects

From Analytics to Cognition: Taking Genomic Science to the Bedside
The goal of this project is to generate actionable intelligence using smart analytics to integrate big-data in the form of omics (genomics, transcriptomics, metabolomics, etc.), clinical data, and longitudinal data from electronic health records (EHR). The actionable intelligence is a descriptive piece of information with high confidence and accuracy that can be used to tailor and individualize diagnosis and therapeutics for a given patient or inform potential candidates for biomarker discovery. The analytics and tools will be developed using engineering expertise at the Univ. of Illinois in close collaboration and partnership with leading clinicians, biologists, and bioinformatics specialists at Mayo Clinic. This project is exploring societally relevant, prevalent, and yet less-understood diseases such as triple-negative breast cancer, major depressive disorder, and diabetes.

Information-Compression Algorithms for Genomic Data Storage and Transfer
Data compression is crucial for enabling timely exchange and long-term storage of heterogeneous biological and clinical data. To facilitate efficient organization and maintenance of genomic databases and to allow for fast random access, query, and search, specialized software solutions for compression and computing in the compressive domain is being developed.

Improving the Accuracy of Genomic Variant Calling Through Deep Learning
This project will develop new deep learning approaches to tackle unsolved problems for variant calling (e.g., SNPs and small indels in low-complexity regions with ambiguity). Unlike traditional methods, our algorithms will not only provide the best variant calling quality but also translate well across different application domains (germline/somatic), sequencing methods (WGS/exome/amplicon), and platforms (Illumina/IonTorrent). Meanwhile, our new machine-learning-based implementation will use industry-standard libraries, such as Tensorflow and STL, and target both GPUs and FPGAs for computation acceleration.

Scaling the Computation of Epistatic Interactions in GWAS Data
Calculating epistatic interactions between genomic variants in studies incorporating complex endophenotypes is a computationally challenging problem that requires emphasis on accelerating and parallelizing the code and achieving workload distribution efficiency. Development of fast production-grade software in this area will enable the detection of epistasis in many existing GWAS datasets, in both the biomedical and agricultural fields.

Proposals and Pre-Proposals for IAB Feedback and Discussion

Connecting genes to phenotypes using a multiscale modeling framework.
Matthew Turk, Amy Marshall-Colon, Meagan Lang, Stephen P. Long (UIUC).
Crop production contributes approximately $56.7 billion annually to the Illinois economy, and is responsible for nearly 200,000 jobs. Climate change models predict historically unprecedented warming by the end of this century, which will result in a temperature-related yield loss of 15% for corn and soybean, or an $8.5 B economic loss over the next 25 years. Illinois needs to develop strategies to mitigate the negative consequences of climate change on crop production. Computational biology and molecular breeding technologies are effective approaches for generating crops that are highly productive under challenging environmental conditions, such as heat, water, and nutrient stress. The scientific challenge is to identify appropriate genetic targets for crop improvement. The computational challenge is to integrate data from genomics and data-informed procedural modeling of crop structure and form to identify meaningful links between genes and phenotypic traits of interest. Computation can help identify important gene candidates for bioengineering to meet the rapidly expanding need for food; however, without new computational tools to enable modeling and visualization of multiscale processes, direct connection between the input parameters and the resultant crop production can be impossible to establish.

In the proposed project we will use the existing Crops in silico framework, cis_interface (https://github.com/cropsinsilico/cis_interface/), to integrate multi-omics data into whole crop models of corn and soybean. The framework enables model integration using asynchronous communication via programming language-neutral API calls, enabling a diverse set of computational models to communicate with minimal change. Within this framework, we link information from the gene-level to the whole-plant and canopy-level, utilizing procedural ("L-system" and ray-tracing) models to construct plant structure and photosynthetic models informed by gene expression. This coupling of architectural models with gene expression, via process-based models of photosynthesis and nutrient and water uptake, provides a means to link genes with structural phenotypes. Whole-crop models will be mechanistically informed by gene expression and metabolic flux models. In a proof-of-concept experiment, we have successfully linked gene expression, flux analysis, and leaf-level photosynthesis models to produce a mechanistically informed multi-omic model of soybean leaf response to changes in atmospheric carbon dioxide. This model predicts crop response to future climatic conditions.

This project is of interest to industrial partners in both biology and computation. Cis_interface can easily be used by modelers and domain experts in any biological field. The biological insights obtained from predictive, multi-scale modeling can identify genes that contribute to economically important crops traits. Computational industry partners like SLS, IBM, and Xilinx could provide both computer science expertise and computational resources for the project and also benefit from the framework, which is designed for integration in domains beyond crop sciences. Intel, with its detailed experience in software for ray-tracing, would be an excellent partner for development of the photosynthetic models. Likewise, the project goals would contribute to the mission and interests of a biological partner like Dow AgroSciences by aligning with their sustainability commitments and economic interests of predicting genetics that contribute to value-added traits. One outcome of this project is the identification of genetic targets for engineering, and thus has the potential to generate IP.

Algorithms for Experimental Study Design in Cancer Genomics.
Mohammed El-Kebir, Sanmi Koyejo (UIUC) and Nicholas Chia (Mayo).
The key challenge in cancer phylogenetics is to infer a phylogeny given sequenced biopsies from the same tumor. Typically, each bulk sample is composed of DNA sequences from millions of cells with distinct genomes. Importantly, the phylogeny inference problem from bulk samples exhibits non-uniqueness of solutions. That is, multiple phylogenetic trees may explain the same input data, potentially leading to divergent conclusions in downstream analyses. Methods for tumor sequencing study design aimed at reducing ambiguity do not currently exist. Here, we propose the first computational method, which, given preliminary sequencing data, will suggest follow-up sequencing experiments with the aim of reducing non-uniqueness of solutions. Our method will be based on a mathematical model that incorporates a tradeoff between non-uniqueness and costs of different sequencing technologies such as (synthetic) long-read, single-cell and targeted sequencing. The proposed method will lead to better sequencing experiments that improve our understanding of tumorigenesis and provide actionable intelligence for personalized medicine. We will validate our method using extensive simulations. We will apply and showcase our method on actual tumors in collaboration with Nicholas Chia (Mayo).

Reverse engineering the human cancer transcriptome.
Mikel Hernaez (IGB, UIUC), Dave Zhao (Statistics, UIUC) and Manish Kohli (Mayo Clinic).
Reverse engineering gene regulatory networks from gene expression data is still a major challenge in computational biology. Gene regulatory networks provide a concise representation of the transcriptional regulatory landscape of the cell, and although they do not reflect post-transcriptional modifications, they have been successfully used in many applications to shed light on new biological mechanisms and gene level relationships in cells. Elucidating the changes in these gene regulatory mechanisms across different conditions (e.g., drug responders vs. non-responders) is fundamental for drug development and personalized treatment. The proposed project aims at building an automatized end-to-end solution for the discovery of biomarkers associated with transcriptional network perturbations underlying with disease and drug response. For example, a mutation in the genome could cause the re-wiring of a particular biological pathway that may lead to greater responsiveness to a drug. It is critical to not only discover the association between the genomic variant and the drug response, but also to understand the mechanistic changes in the transcriptome that the variant is causing. The discovered associations will be further integrated with existing clinical data to deliver actionable intelligence for physicians to elucidate the best drug treatment for the patients at hand. Furthermore, the identified biomarkers like SNPs and perturbed pathways could be used for the development of tailored drugs. The proposed pipeline will be tested on data provided by Mayo Clinic on metastatic castrate-resistance prostate cancer (mCRPC). The aim of this particular application is to find the biomarkers (both genomic variants and perturbed biological pathways) that associate with the response of mCRPC patients to the Abiraterone drug, one of the most widely used drugs for this type

of cancer. By discovering such biomarkers, we expect to help physicians provide a personalized selection of the best treatment for each patient.

Accelerating multiple hypothesis testing through GPU/FPGA hardware accelerators and quantum computing infrastructure: Application to genome-wide association studies.
Eric Jakobsson, Professor Emeritus UIUC, NCSA Associate; Volodymyr (Vlad) Kindratenko, Associate Professor UIUC, NCSA Senior Research Scientist; Alexander E. Lipka, Assistant Professor UIUC.
The need to perform many instantiations of multiple hypothesis testing is a problem in analyzing complex systems that would immensely benefit from specialized hardware technologies. We propose to implement massive multiple hypothesis testing in two novel ways, one utilizing existing GPU/FPGA hardware accelerators to speed up costly segments of the process and another one that harnesses existing and accessible quantum computing architectures by mapping the multiple hypothesis problem to known quantum algorithms and executing test data.

Innovative Systems Laboratory (ISL) at NCSA will provide the required development expertise and environment, including an NVIDIA V100 GPU system and Xilinx Kintex UltraScale FPGA KCU1500 Accelerator Both of these systems are state-of-the art hardware for accelerating computationally intensive tasks. We will study and evaluate suitability of IBM Q platform and Intel's latest quantum computing chip .

Overall, we anticipate our tools to be of interest to technology partners, applicable to realistic bioscience challenges faced by Mayo partners, and relevant to identifying drug targets of potential interest to pharmaceutical partners. As a testbed problem, we will choose analysis of GWAS data.

Design and Assessment of Secure Genomic Pipeline
Zbigniew Kalbarczyk, UIUC.
This project addresses a fundamental need for clinicians and researchers/computer scientists to work together to advance precision medicine. The foundation of this collaboration between clinicians and researchers is the ability to share electronic health data and patient specific biological data. However, this presents a huge challenge that involves maintaining security and trust of many points 1) the individuals accessing the data; 2) devices and endpoints connecting to networks where the data resides; 3) data residing in permanent or temporary storage; and 4) data in computation/analysis.

This project will study genomic data lineage that includes data's origin and its flows overtime; simulate authentication-based attacks targeting data flows from the attacker's point of view; and develop detection technique to correlate pre-attack activities (e.g., scans and/or accessing remote servers) from both host and network monitors to preemptively stop the attack.

GENOMIC DATA COMPRESSION

Center/Site: Center for Computational Biotechnology and Genomic Medicine (CCBGM)
Tracking Number: 1.1.2
Project Leader(s): Olgica Milenkovic, UIUC
Email: milenkov@illinois.edu
Type: Continuing
Proposed Budget/Years: $60K
Faculty Collaborator(s): Idoia Ochoa, UIUC
Start Date: January 2017
Estimated Project Completion Date: August 2019

Project Description/Overview:
The project is focused on developing novel lossless and lossy compression methods and supporting machine learning tools for large-scale multiomics data. In particular, the work so far has focused on developing compression methods for raw sequencing data, reference-based metagenomics data storage, methylation data analysis and lossless compression, and more recently, on scRNA-seq data clustering for compression.

Progress to Date:
To date, we finished three lines of work, pertaining to compression and quantization of RNA-seq, ChIP-seq data, and methylation data. In the context of methylation data compression, we also developed the first quantized deep learning methylation pattern prediction platform, and applied it to pan-cancer datasets available from TCGA.

The team members would like to collaborate with industry partners on FPGA implementations of some of the existing and new compression methods for scRNA-seq data.

What has changed since the initial plan, include any changes in assumptions.
The only change in the original project is the inclusion of new machine learning tools into the compression platform and design of new learning tools that facilitate statistical data analysis and hence more efficient compression.

Timeline – Milestones and Deliverables:
We have started with the scRNA-seq data compression and are currently developing multivew nonnegative factorization (NMF) methods that may be used for data clustering and consequently reference-based on group-based compression of transcriptome data. We expect to have preliminary results in the next 3-4 months. Part of this work is also supported by the Chan-Zuckerberg Human Cell Atlas award to PI Milenkovic.

Experimental plan (current year only and what has changed from original plan):
No changes are made in the original plan. We are planning to use publicly available scRNA-seq data as well as some of the data made available to us by the Chan-Zuckerberg initiative to test our new, distributed NMF algorithms.

Summary of research accomplishments since last meeting:
A paper on the topic of pan-cancer analysis of mutual information between RNA-seq and methylation profiles is currently under preparation and should be available for review by the industrial partners within a month.

Spotlight on Students:
Jianhao Peng is supported by the grant, and the primary student investigator on problems pertaining to methylation data analysis and scRNA-seq data compression.

Potential Member Company Benefits:
scRNA-seq data is one of the fastest growing genomic data source and scRNA has the potential to reveal many previously unknown regulatory mechanisms in cell. Being at the forefront of scRNA data storage system design provides unique opportunities for life science research and infrastructure management.

FROM ANALYTICS TO COGNITION: TAKING GENOMIC AND OTHER "MULTI-OMIC" SCIENCE TO THE BEDSIDE

Center/Site: Center for Computational Biotechnology and Genomic Medicine (CCBGM)
Tracking Number: 1.1.5
Project Leader(s): Ravishankar Iyer (UIUC); Liewei Wang and Richard Weinshilboum (Mayo).
Email:
Type: Continuing
Proposed Budget/Years: $
Faculty Collaborator(s): Deming Chen, Wen-mei Hwu, Zbigniew Kalbarczyk (UIUC); Jim Ingle, Matthew Goetz, Krishna Kalari, Mark Frye and William Bobo (Mayo).
Start Date: January 1, 2017
Estimated Project Completion Date: November 2018

Project Description/Overview:
The increasing volume of human genomic sequencing and other "multi-omics" data, along with the growing ability to analyze such data quickly and efficiently together with rich data from electronic health records (EHR), provide new opportunities for innovation in healthcare research, science and delivery. Of particular clinical interest is the ability of rich datasets and methodological innovations that can enhance our understanding of disease and its treatment and, ultimately, improve patient outcomes. Such improvement in patient outcomes can be achieved by accurate mapping of patients to drugs using biologically-guided actionable intelligence derived from statistical/machine learning techniques. A major challenge in generating such actionable intelligence is how exactly longitudinal data (both patient-specific and population-based) obtained from EHR will be integrated with omics data to improve predictability in treatment outcomes. In this project, we aim to predict the long-term treatment outcome in patients with breast cancer treated with aromatase inhibitors. The outcome of interest is "tumor recurrence", which causes over 40,000 deaths in the US alone. To achieve this goal, data from over 1,500 patients will span hormone levels (pre-, and post-treatment), EHR, genome-wide genomics and clinical measures. Guiding the analyses of this project is our prior experience in predicting antidepressant treatment outcomes in patients with major depressive disorder and breast cancer, wherein methods were developed to integrate genomics, metabolomics, clinical assessment of depression severity and social/demographic factors to generate clinically actionable intelligence.

Progress to Date:
Using data from triple-negative breast cancer (TNBC) and major depressive disorder (MDD) as drivers, we have demonstrated the ability of,

1. Data-driven model-based unsupervised learning to identify a sub-population of single-cells (TNBC) which are differentially expressed after metformin treatment.
2. Machine learning workflows with probabilistic graphs as the core of the methods, which by combining patient-specific omics data with social/demographic information identifies a subset of MDD patients who achieve early remission vs. non-response to antidepressant treatment.

Overarching significance of this work is the ability of analytical methods to help better understand therapeutic efficacy of drugs by identifying novel biomarkers of drug response. These biomarkers are the candidates for laboratory verification which explains patient-specific variation in drug response. Thus, the methods are validated for further extensive clinical validation – thereby highlighting the value in close collaboration with Mayo and UIUC teams.

What has changed since the initial plan, include any changes in assumptions.
The project's original goal was aimed at predicting treatment response in MDD patients treated with antidepressants. We are now broadening the breadth of the disease we study by also studying long-term treatment outcome in breast cancer – a major disease prevalent largely in women. Specifically, the goal is to predict long-term breast tumor recurrence in patients treated with aromatase inhibitors.

Timeline – Milestones and Deliverables:
Year 1 (Nov '16 to Nov '17): Using data from Mayo Clinic-PGRN AMPS, STAR*D and ISPC datasets, we established predictability in antidepressant treatment outcomes in patients with MDD. Methodological innovations comprised 1) using mixture-model based unsupervised learning to identify patient stratification at all time-points of the trial based on their depression severity scores – allowing for an overlay of associated metabolomics and genomics data, 2) using probabilistic graphs that allow for study most-likely longitudinal variation of depressive symptoms in patients stratified by their depression severity during the trial, and 3) the use of supervised learning methods to predict treatment outcomes in patients by combining social/demographic measures (from EHR), biological measures and depression severity measures. Through these innovations, we demonstrated that sex-specific predictions in antidepressant outcomes can be achieved with AUC > 0.7, and with cross-trial replications.

This work was published at the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), which was recipient of Best Paper Award. This work also received the American Society of Clinical Pharmacology and Therapeutics' Presidential Trainee Award in 2017 and 2018 and Jason Morrow Presidential Trainee Award in 2017.

Project Progress (Nov '17 – May '18):
We have now pooled and prepared the datasets of over 1,500 subjects treated with anastrozole (an aromatase inhibitor) from the NIH funded MA-27 study and the joint study of Mayo Clinic, Memorial Sloan Kettering Cancer Center and MD Anderson Cancer Center. This data comprises hormone levels (pre-treatment, and post-treatment), genome-wide genomics and clinical measures (for e.g., BMI, tumor staging, tumor size, prior treatments). The goal is to use similar approach used in methods developed to further the understanding antidepressant response and mechanisms in MDD patients, to

1. Subtype of patients who present unique characteristics of tumor recurrence.
2. Identify biological characteristics in hormones and genome that differentiate eventual tumor recurrence.
3. Predict eventual tumor recurrence within 5 years of treatment completion (i.e., 6 months of anastrozole treatment) in patients based on their baseline and post treatment characteristics.

Experimental plan (current year only and what has changed from original plan):
Though continued interaction with clinicians, the clinical impact of developed methods is multi-fold when they are rendered in the form of a website which can be interactive with physicians and explain the results of the analyses. Thus, we have focused this past quarter on web-tool development in addition to developing the proposed analytics.

Summary of research accomplishments since last meeting:
A paper on the topic of pan-cancer analysis of mutual information between RNA-seq and methylation profiles is currently under preparation and should be available for review by the industrial partners within a month.

Spotlight on Students:
UIUC undergraduate students Mahima Desetty and Connor Aubry have developed a working web-interface for psychiatrists to enter patient information, which collected by the developed analytics can predict treatment outcomes. Graduate student Arjun Athreya has accepted a full-time position at Mayo Clinic as an Associate Consultant-II and Assistant Professor. Expected to join the clinic in Fall.

Potential Member Company Benefits:
While methods have been developed with MDD and breast cancer as drivers of analyses, we believe that the methods are broadly applicable for studying

1. Therapeutic efficacy of other drugs in other diseases (of interest to CCBGM's pharmaceutical IAB members)
2. Causes for early response to treatment beyond human diseases (of interest to CCBGM's agro-science IAB members)
3. Integration of tools in broader healthcare technology & service industry (of interest to CCBGM's technology services IAB members)
4. Standardization of methods which could be implemented in hardware/software accelerators to chieve performance gains on cloud infrastructures (of interest to CCBGM's computing technologyl IAB members)

IMPROVING VARIANT CALLING THROUGH DEEP LEARNING

Center/Site: Center for Computational Biotechnology and Genomic Medicine (CCBGM)
Tracking Number: 1.2.2
Project Leader(s): Deming Chen, Steven S. Lumetta (UIUC), Eric Klee (Mayo Clinic)
Email: dchen@illinois.edu
Type: Continuing
Proposed Budget/Years: $150,000 including 10% overhead
Faculty Collaborator(s): Liudmila Mainzer, Zbigniew Kalbarczyk, Olgica Milenkovic, Sandra Rodriguez-Zas, Matthew Hudson (UIUC); Jean-Pierre Kocher, Eric Wieben, George Vasmatzis (Mayo)
Start Date: August 28, 2017
Estimated Project Completion Date: August 27, 2020

Project Description/Overview:
Variant calling using Next Generation Sequencing (NGS) technology has wide range of applicability in modern bioinformatics and genomics. Discovering the underlying genetic traits of Mendelian diseases (Bamshad et al., Nature Genetics 2011), analysis of oncogenes in cancer genomics (Yang et al., BMC Medical Genetics 2010) and study of genetic diversity to help strategize crop-breeding methods (Jiao et al, Nature Genetics 2012) are just a few of these applications.

Single Nucleotide Polymorphisms (SNPs) and small indels, which account for the vast majority of mutations in a typical human genome (The 1000 Genomes Project consortium, Nature 2015), are of great interest in many of these studies. Though there are many mature pipelines for discovering these types of mutations in various scenarios (GATK for germline, VarScan2 for somatic mutations, Stacks for GBS-based studies), there are still many unsolved problems in this area. For example, many popular tools have been found to consistently underperform in certain areas of the genome for germline SNP and indel calls for human whole genome data (Li, Bioinformatics 2014). The effect will only be amplified in more complex settings such as somatic mutation calls or where the allele fraction is lower. Such imperfections arise from the use of hand-tuned pre- and post-processing steps which are sub-optimal, as well as simplifying assumptions adopted by the tools which are not representative of errors and artifacts in sequencing and read alignment. In addition, these methods do not translate well across sequencing technologies and error models, and there is no systematic method to tune the parameters of the underlying algorithms to suit specific needs.

Deep Neural Networks (DNNs) have shown great promise in complex classification tasks. For example, during the training phase, a Convolutional Neural Network (CNN) can learn the type of features needed for image classification as well as tune the filters needed to discover these features in the input data, eliminating the need for expert-driven image pre-processing steps, outperforming conventional methods (Krizhevsky et al, NIPS 2012). A CNN-based approach to variant calling, DeepVariant (Poplin et al., bioRxiv 2016), which treats the pileup data from read alignment as images has been shown to be highly successful in germline SNP/indel calls.

We propose to specialize the deep learning approach to variant calling for SNPs and small indels, targeting difficult-to-call regions of the genome. However, instead of adapting variant calling to another problem like image classification, we propose to develop models that are native to sequence representation and read alignment within a deep learning framework to obtain a more faithful and complete formulation of the problem. This approach will allow the characteristics of ground truth variants in a training set to be reflected across the entire pipeline, starting from bam preparation steps (e.g., realignment) to the generation of the final call set. Unlike traditional methods, our algorithms have the potential to translate well across different application domains (germline/somatic), sequencing methods (WGS/exome/amplicon), and platforms (Illumina/IonTorrent), since the core computations use sophisticated models which can be adapted to each situation systematically using the well-known training methods for neural networks. We plan to generate datasets for training our deep network solution in a variety of these cases and provide trained models which may be directly deployed for variant discovery. We will also provide necessary software to train the model with a user-provided dataset in case a laboratory or clinic wants to adapt the tools to a specific sequencing pipeline.

This overarching objective can be divided into two parts. The first is the creation of a deep learning architecture that can analyze a site of interest and arrive at a decision as to whether the site contains a variation. One novel component of the architecture will use unsupervised and semi-supervised learning to produce representations of sequences and sequence pairs that expose the relationship between read and reference at a greater level of detail than the traditional pileup representation. This representation will be in a format that may be consumed by well-known DNN layers such as convolutional layers. Since this component replaces the alignment step (localized) performed during variant calling, we call this the alignment layer. While the alignment layer can summarize the relationship between read and reference well, it does not explicitly model the probabilities of misalignment where a read is mapped to a completely wrong location in the reference. A context-dependent estimate of the probability of mismapping can be a valuable tool in discounting reads representing bad evidence with high specificity, allowing for the reduction of false positive calls. A simplistic estimate of the probability can be obtained from a look-up table which maps reads or reference contexts to mismapping probability. However, the memory requirements of this look-up table can be very large, or it can overfit to a particular reference sequence and release version, whereas it is possible to approximate the look-up table functionality efficiently using a neural network, with reduced risk of overfitting due to memorization. Overall, erroneous alignments cause a large proportion of false calls in low-complexity regions (Li, Bioinformatics 2014), where variant calling has been difficult. The neural network components proposed here directly address this limitation. The components will be initially trained separately to optimize their respective learning objectives (improving sequence-pair representation, and predicting the likelihood of mismapping), and then their outputs will be fed into well-known DNN layers to obtain the variant call. The complete architecture can then be trained end-to-end to improve overall variant call accuracy.

The second part of the project is the creation of a dataset for training the networks. The dataset will consist of training and testing vectors for reads from different sequencing technologies (e.g., Illumina/PacBio), for different use cases (exome/amplicon/WGS) and applications (somatic/germline). The publicly released version of the dataset will contain simulated read sets created from anonymized or publicly available real sequencing data. Variants will be simulated in silico by mixing real reads from multiple samples. Clonal heterogeneity for somatic calls, and ploidy for germline calls will be simulated. The dataset will contain complex and simple regions in the reference, as well as known locations of variations and random locations in the genome to prevent over-fitting. In our internal training pipelines, additionally, we will include real sequenced data and real variants; in this case, the ground truth will be verified from sequencing using two different technologies (e.g., Illumina and Sanger, or Illumina and PacBio).

References

1. Bamshad, et al. (2011), “Exome sequencing as a tool for Mendelian disease gene discovery,” Nature Review Genetics, 12:745-755
2. Yang, et al. (2010), “Important role of indels in somatic mutations of human cancer genes,” BMC Medical Genetics, doi: 10.1186/1471-2350-11-128
3. Jiao, et al. (2012), “Genome-wide genetic changes during modern breeding of Maize,” Nature Genetics, 44:812-815
4. The 1000 Genomes Project Consortium (2015), “A global reference for human genetic variation,” Nature 526:68-74
5. Li (2014), “Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples,” Bioinformatics 30:2843-2851
6. Krizhevsky, et al. (2012), “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in Neural Information Processing Systems 25:1097-1105
7. Poplin, et al. (2016), “Creating a universal SNP and small indel variant caller with deep neural networks,” bioRxiv, doi: https://doi.org/10.1101/092890

Progress to Date:
Creation of the alignment layer and initial network for variant calling.

- A prototype for the alignment layer is ready, that can represent sequence pairs. This is generated based on a Hidden Markov Model (HMM) that enumerates all possible edits of the reference.
- A network training and testing infrastructure has been written. It can create CNNs in a general Directed Acyclic Graph format as specified by the user, accept the alignment layer outputs and provide a variant call. Currently, it has been applied to germline SNV calling (using real data from GIAB) in low-complexity regions, and it is able to call variants at an accuracy comparable to that of GATK.

Pipeline for training/test data generation. This includes two components.

- The first component accepts a real BAM file as input and performs two things. (1) It can spike-in a mixture of randomly chosen and known mutations by altering reads at specific locations to simulate

mutations and (2) stratify locations according to their complexity and provide a training and test set at a given level of complexity; this allows us to concentrate DNN training and testing on challenging areas in the genome.
- The second component accepts a BAM file and applies simple heuristics to provide a first-level variant call that is highly sensitive, but not highly specific. The result is a set of sites in the reference where there is a putative variant with approximately a 50% chance which is friendly for DNN training where the target is one of two labels. This currently works for SNVs.

A novel generative model for sequences that generalizes the HMM, called the Long Short-term Graphical Model (LSGM).
- We combined the highly successful Recurrent Neural Network (RNN) model used for sequence-type data analysis (speech, genomics etc) with the Hidden Markov Model used for many applications in bioinformatics including variant calling (e.g., pair-HMM used in GATK).
- The model can directly replace the HMM in most applications where forward/backward/Viterbi inference is used. The model can be trained to maximize the likelihood of the training set, similar to the HMM. At the same time, the model can potentially represent long-term patterns, where RNNs have outperformed the HMM. So, it combines the inference flexibility of the HMM with the long-term pattern recognition capability of the RNN.
- The model outperformed the HMM and other DNN-based solutions in predicting transcription factor binding specificity in a well-known benchmark consisting of five transcription factors.
- The model outperformed DNN-based methods in two out of four tracks in a well-known music benchmar
- Given that the alignment layer is constructed using an HMM, we see potential to replace the HMM with the LSGM.

Performance optimizations of the HMM. These will allow us to scale to larger training sets. Our current training set is small.
- We have implemented our alignment layer HMM with slight modifications using pytorch that is friendly towards SIMD-based systems. It also allows end-to-end training when embedded into our variant calling framework, unlike the previous implementation. The original implementation used C++ without SIMD constructs but could be run on multi-core CPU. The new implementation shows performance improvements when using a GPU over the multi-core CPU C++ version. The pytorch implementation hasn't been tested on multi-core CPU yet.
- We have done some preliminary experiments to use dilated convolutions to mimic the alignment layer outputs. Initial experiments show some ability to mimic the alignment layer representations. If this can be finalized, we will switch from the HMM to this method, since this can be parallelized much more efficiently on the GPU than the HMM.

What has changed since the initial plan, include any changes in assumptions.
In the initial plan, we envisaged a complete end-to-end variant call pipeline modeled by a monolithic DNN, with the addition of the alignment layer. Initial models beating state-of the-art tools weren't expected to be affected by throughput considerations. This picture has changed in the following ways:
- The alignment layer is one of the components. It is capable of representing read and reference pairs well, given the reference location where a read is mapped. However, it may not be able to learn mismapping probabilities which also have a significant impact on variant call accuracy.
- Mismapping probabilities will be modeled using a separate neural network, based on simulated reads
- Throughput optimization of the alignment layer is important. Without optimization, the pipeline cannot process a sufficiently large training set. We have SIMD-friendly implementation of the HMM in the alignment layer and are also attempting to replace the HMM with dilated convolutions. We also propose to look at Strelka 2's sampling method to replace the HMM.

Timeline – Milestones and Deliverables:
Year 1: Develop the necessary infrastructure (the alignment layer) for constructing the deep neural network for variant calling. Construction of a shallow network for initial training, testing. Generation of training and validation pipelines for Illumina datasets.

Year 2: Model mismapping probabilities using a neural network. Optimize alignment layer throughput for scaling to larger training sets. Initial explorations with TGS reads. Complete a deep-network implementation of variant calling flow. Study network performance side-by-side with more complex data. Study the balance among various types of training vectors in the data – random/known variations, simple/difficult to call sites. Enhance benchmarks for comprehensive context-specific metrics for goodness of call.

Year 3: Use knowledge from year 2 to generate the complete suite of datasets for exome/amplicon/WG sequencing for somatic/germline calls. From Mayo's clinical and research pipelines, extract sites that are verified using two different sequencing platforms, to include in internal training and validation. Train different instances of deep network for variant calling in different scenarios. Deliver open-source code for training, and inference, and open source models for direct deployment.

Experimental plan (current year only and what has changed from original plan):
- Improve alignment layer throughput performance, and scale training to larger datasets. Complete throughput evaluation using SIMD-friendly implementation and integrate into existing flow. Complete the study of using dilated convolutions to replace the HMM in the alignment layer.
- Model mismapping probabilities using neural networks. Different tools model mismapping using different methods such as heuristics, and we plan to replace them with our method. We will try to quantify the improvement using our method by inserting this technique into existing tool flows.
- Data analysis using various training/test sets from our data preparation scripts. Examine use of TGS read databases using our data preparation scripts (e.g., PacBio from GIAB).

Summary of research accomplishments since last meeting:
- Completed rebuttal process for NIPS 2018 for the LSGM paper. Improved performance of the model in music modeling.
- Implemented SIMD-friendly version of HMM using pytorch.
- Anand Ramachandran, student working on the project, interned at Illumina gaining valuable experience that will help the project going forward.

Spotlight on Students:
Anand Ramachandran is a PhD student working on the project.
Huiren Li who has been working on the project, graduated last semester with a Master's degree.

Potential Member Company Benefits:
Industrial partners can expect a variant-calling pipeline that is portable across various sequencing platforms and settings, giving higher accuracy than state-of-the-art methods. Our new machine-learning based implementation will use industry-standard libraries such as pytorch, Tensorflow and STL, and the deployment of our algorithms on a variety of computing platforms will showcase the performance, usability, and flexibility of new high-performance hardware and tools.

SCALING THE COMPUTATION OF EPISTATIC INTERACTIONS IN GWAS DATA

Center/Site: Center for Computational Biotechnology and Genomic Medicine (CCBGM)
Tracking Number: 1.2.3
Project Leader(s): Liudmila Sergeevna Mainzer (U. of Illinois)
Alexander Edward Lipka (U. of Illinois)
Nilufer Ertekin-Taner (Mayo Florida)
Liewei Wang (Mayo Rochester)
Email(s): lmainzer@illinois.edu, alipka@illinois.edu, taner.nulifer@mayo.edu, wang.leiwei@mayo.edu
Type: Continuing
Proposed Budget/Years: $120,000 / 2 years
Faculty Collaborator(s): Eric Jakobsson (U. of Illinois); Curtis Younkin, Mariet Allen (Mayo Florida); Inna Ovsyannikova, Richard Kennedy (Mayo Rochester)
Start Date: June 1, 2017
Estimated Project Completion Date: June 1, 2019

Project Description/Overview:
The goals are (1) to design a scalable, efficient and easily deployable software for GWAS on complex quantitative traits, enabling derivation of a complete model for additive and epistatic interactions of multiple orders; and (2) to experiment with statistical approaches to reduce or circumvent the multiple testing problem that arises in epistasis analyses on very large genotypic datasets.

Progress to Date:
Fall 2017: Constructed several code prototypes for efficiency of algorithm implementation and data structures. SPAEML was implemented in Scala for Apache-Spark using DataFrames as well as using RDDs; experimented with Hadoop MapReduce, Neural Networks using TensorFlow and LASSO in Scala for Apache-Spark.

Spring 2018: Confirmed that SPAEML outperforms traditional Joint Linkage, single-locus and single-pair epistatic analyses in terms of accuracy, for both maize inflorescence and human AD traits. Developed script prototypes to pull SNP annotations from multiple databases and rank them according to proximity to various genomic features in the effort to prioritize them for inclusion into the final model.

We need access to larger datasets to do scalability analyses and test performance improvements. At the moment we only have access to public GWAS data with a few thousand individuals. We need a dataset with a few tens of thousands of individuals. Perhaps the Industry members can assist. Happy to sign NDAs.

What has changed since the initial plan, include any changes in assumptions.
This Summer 2018 we obtained a substantial allocation of Amazon Research Credits in AWS, and also a startup allocation on XSEDE, to scale-out the EpiQuant software across more nodes and improve the runtime.

Timeline – Milestones and Deliverables:
Fall 2018: Will improve code robustness for SPAEML in Apache Spark, deploy in AWS and begin scalability experiments. Will apply for funding to support this work further. Candidate RFAs: PA-18-867, NSF-18-566, PAR-18-843, NSF-18-567.

Spring 2018: Will focus on automation of search space reduction methodologies.

Experimental plan (current year only and what has changed from original plan):
In the Fall 2018 and Spring 2019 we will automate the scripts for SNP annotation and prioritization and refocus on designing the workflow that ties together LASSO, SSR, SPAEML and biased Monte-Carlo, to progressively refine the multilocus model of additive and epistatic effects. We are targeting the Fall as the submission timeline for an NIH proposal to support further software development in this project. The next goal, for Spring of 2019, is to convert this workflow into a containerized solution for deployment on various compute infrastructure.

FUNDED PROJECTS

Summary of research accomplishments since last meeting:
Manuscript entitled “An assessment of true and false positive detection rates of stepwise epistatic model selection as a function of sample size and number of markers” was submitted to Heredity in early July, 2018.

Spotlight on Students:
Weihao Ge was an integral part of the project in Year 1, and recently defended her PhD thesis. She now accepted a post-doctoral fellowship with us, to continue the work on the project.

Potential Member Company Benefits:
Most GWAS data collected by academia or industry have been analyzed only for associations between phenotypes and polymorphisms at single loci. It is well known that these associations account for only a small proportion of the known heritability of the phenotypes, and that full analysis of potential epistatic effects would immensely increase the usefulness of the data. The procedures and software developed by the project will immediately benefit research in neurodegenerative diseases, immunology, and agronomically important maize traits, due to the involvement of the respective co-PIs and collaborators. However, those products will be equally applicable to other complex traits of medical and agricultural import, such as obesity, autism, and plant disease resistance. Cloud deployment of the scalable software will further facilitate adoption in areas where multidimensional datasets comprised of transcriptomics and metabolomics, as well as in vitro or in vivo phenotypes, pose a severe computational challenge.

INDUSTRY MEMBERSHIP

The real value of the Center is its breadth in approaching the big data problem, from analytics to actionable intelligence, in a comprehensive manner. The Center spans biological expertise ranging from human genomics to crop and animal sciences, and closely melds this with broad expertise in computing systems and algorithms (from HPC to the cloud and special-purpose acceleration). Members will get the benefit of working with industrial and academic partners in the domains of computing, biotechnology, and health science.

Through this process, they will have access to research teams composed of students, researchers, and faculty in computing systems, bioinformatics, genomic applications, and health from Mayo Clinic and the University of Illinois.

Benefits to Industry Members Include:

- Highly leveraged and cost-effective research with overhead fixed at 10%
- Nonexclusive royalty-free license to intellectual property
- Seat on Industry Advisory Board that votes on the direction of Center research and projects
- Recruiting opportunities for graduate students and postdocs
- Access to students involved in Center research
- Access to research teams in computing systems, bioinformatics, genomic applications, and health from the major research university partners (Mayo Clinic, University of Illinois at Urbana-Champaign)
- Opportunities to collaborate on NSF proposals, including Collaborative Research Between I/UCRCs (CORBI), Accelerating Innovation Research (AIR), and Fundamental Research Program (FRP) initiatives

Center Facilities and Resources Available to Members:

At the University of Illinois at Urbana-Champaign

- Beckman Institute for Advanced Science and Technology (BI)
- Carl R. Woese Institute for Genomic Biology (IGB)
- Coordinated Science Laboratory (CSL)
- HPCBio
- CompGen
- Sequencing Unit of the Carver Biotechnology Center
- Merged Computer Infrastructure of HPCBio, Carl R. Woese Institute for Genomic Biology, and Carver Biotechnology Center

At Mayo Clinic

- Organization of Mayo Medical Center
- Center for Individualized Medicine
- Next Generation DNA Sequencing
- Bioinformatics Program and Service Lines
- Information Technology Program

CENTER PROPRIETARY

INDUSTRY MEMBERSHIP

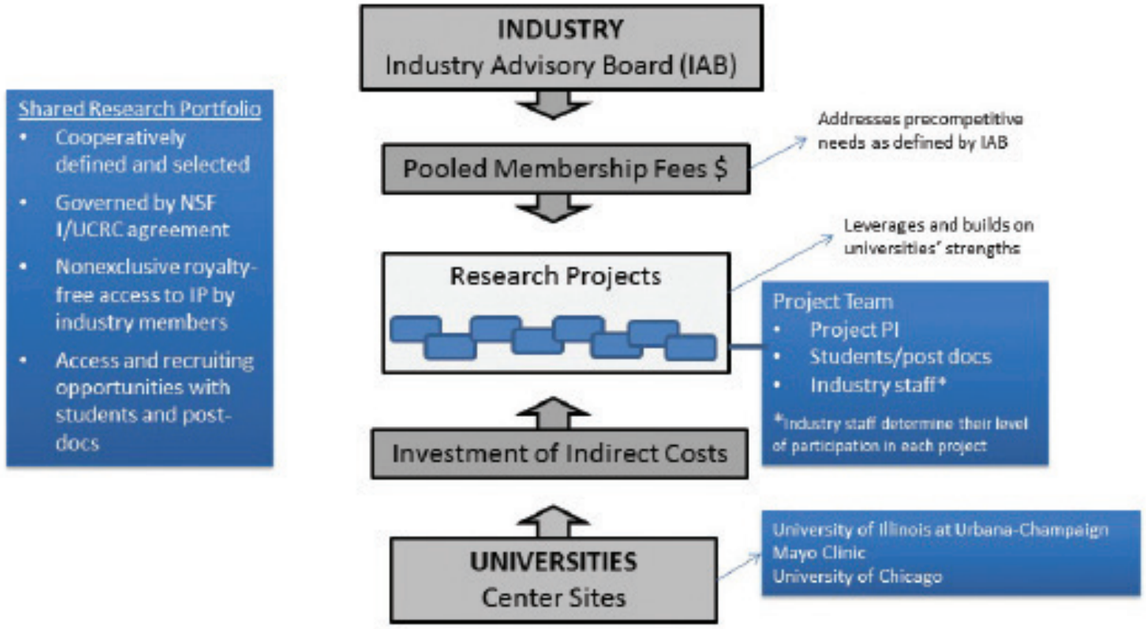
Annual Membership Fee

Membership fees are the primary funding mechanism for research, development, and education undertaken by the Center. The NSF provides some direct funding to each university for up to eleven years, and offers the potential of additional support for specific endeavors. The Center is expected to be sustained by industry. Industry members pay the Center an annual membership fee for each Industrial Advisory Board seat.

Industry Advisory Board (IAB) Meetings

An Industry Advisory Board (IAB) has been established to facilitate industry partnerships and to advise the Center on all aspects of operations, including strategic planning and research assessment. One of the primary roles of the IAB is to help ensure that the research being conducted is in line with industry objectives. IAB meetings are held twice a year to share information about the research conducted at the universities and to collect feedback and input on research needs from industry.

Overview: Center for Computational Biotechnology and Genomic Medicine



INDUSTRY ADVISORY BOARD (IAB) MEMBERS

(AS OF SEPTEMEBER 2018)

Abbott Molecular

Dow AgroSciences

Infosys

Sandia National Laboratory

IBM Systems

Intel Corporation

OSF Healthcare

Xilinx, Inc.

For more information or to become a member, contact:

Ravi Iyer, Center Director, University of Illinois, rkiyer@illinois.edu
Liewei Wang, Center Co-Director, Mayo Clinic, Wang.Liewei@mayo.edu
Kathleen Atchley, Center Associate Director, University of Illinois, katchley@illinois.edu
Leila A. Jones, Research Operations Manager, Mayo Clinic, jones.leila@mayo.edu

NOTES

CCBGM
BIANNUAL MEETING
SEPTEMBER 10-11, 2018

The Mayo Civic Center
30 Civic Center Dr SE | Rochester, MN 55904

CENTER FOR COMPUTATIONAL BIOTECHNOLOGY AND GENOMIC MEDICINE (CCBGM)