National Center for Integrative Biomedical Informatics

University of Michigan

National Center for Biomedical Informatics (NCIBI)

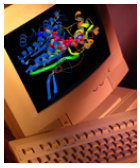
http://www.ncibi.org

The mission of the National Center for Integrative Biomedical Informatics (NCIBI), founded in October 2005, is to create targeted knowledge environments for molecular biomedical research that help guide experiments and enable new insights from the analysis of complex diseases. NCIBI develops efficient software tools, data integration methods, and systems modeling environments. The resulting NICIBI "Suite of Tools and Data" facilitates rapid construction of context-appropriate molecular biology information schemas from experimental data, biomedical databases, and the published literature.

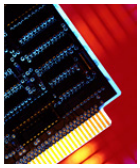
The NCIBI supports information access and data analysis workflow of collaborating biomedical researchers, enabling them to build computational and knowledge models of biological systems validated through focused work on specific diseases. The initial Driving Biological Problems are:

- Prostate cancer progression
- Organ-specific complications of type 1 diabetes
- Genetic and metabolic heterogeneity of type 2 diabetes
- Genetic susceptibility and phenotypic subclassification of bipolar depressive disease

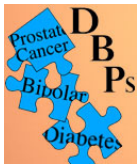
The NCIBI is composed of seven tightly integrated cores. Interactions among cores are extensive and include shared personnel, shared computational infrastructure, a multitude of collaborative research projects, and joint participation in multiple training programs including Bioinformatics, Computer Science, Information Science, Biostatistics, Genetics, Psychiatry, Diabetes and Cancer Biology. The Cores that compose the Center are as follows:



Core 1: Computer and Information Science core, is advancing biomedical computing capabilities for flexibility, scalability, extended lifetime, and navigation through highly complex data and information from literature. This core is developing a computing architecture that permits the harnessing of diverse data sources and algorithms into a single cohesive framework for problem solving by biomedical scientists, a user interaction architecture, user interfaces, querying mechanisms, data analysis algorithms, and techniques to deeply integrate data from multiple sources.



Core 2: Bioinformatics and Knowledge Acquisition core, is building an integrated information store for molecular biology and software tools for information retrieval, bioinformatics analysis, model construction and model testing. The Observational Data Repository (ODR) is structured around experimental data while the Integrated Knowledge Base (IKB) is organized by models for biological processes. This core is also developing software tools for exploratory bioinformatics analyses, model construction and model evaluation software linked to the ODR and IKB

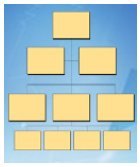


Core 3: Researchers from the first Driving Biological Problems (DBP) core are solving problems from biological domains of Prostate Cancer, Diabetes Type I, Diabetes Type II, and Bipolar Disorder in the NCIBI.

Core 4: Infrastructure core provides the Computational Infrastructure for the Center including hardware management, user authentication, and network communications.

Core 5: Education core is establishing innovative education and training programs, including a new Certificate in Integrative Biomedical Informatics that will leverage the Center's integrated information and computational resources.

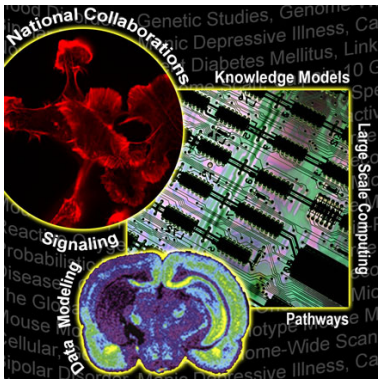
Core 6: Outreach and Dissemination core is aiding in developing problem solving techniques in the collaborative study of complex biological systems. It is building collaboratory web-portal knowledge environments and disseminate Center software technology and data generated in the NCIBI.



Core 7: Administration core provides the administrative management of the Center including an NIH External Advisory Committee, a Software Dissemination, IP, and Data Sharing Committee, IRB compliance, and the NCIBI Executive Committee responsible for overseeing active DBPs, recruiting applications for new DBPs, and reviewing these applications.

Core 1: Computer and Information Science

The central challenge driving this core is the vast number and wide variety of data sources of value to the biomedical scientist—not just public databases, but also collections of experimental observations, information extracted from the literature, and so forth. The set of tools to analyze these data sets is also correspondingly varied. Our mission is to subsume all this complexity and provide the biomedical scientist access to this information through an intuitive and simple interface.



Given the wide variety in the data sources and the types of questions asked of them, it is unreasonable to expect a "one size fits all" solution to this daunting problem. Rather, we are finding that several integration techniques, storage models, and query interfaces are required, each being superior to the others in some contexts. As such, we are pursuing a multiple and diverse number of these, and developing.

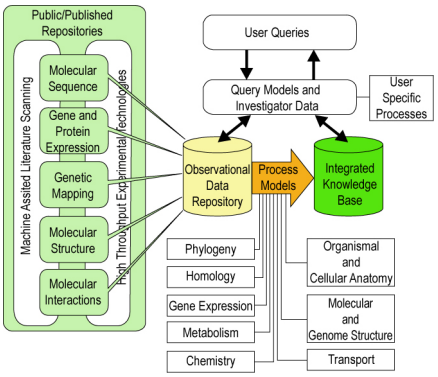
Biological systems are complex, and in many cases our understanding of these systems is less than perfect. In our quest to simplify the information access interface, it is critical that we not simplify away useful information. To this end, we are developing rich data and knowledge models capable of representing all of the biological complexity required. In particular, we are, in contrast with commercial databases, explicitly providing for the representation and management of uncertainty in the database. Furthermore, we are tracking each piece of information with its provenance, through multiple derivation steps if necessary.

The Computer and Information Science Core focuses on developing novel computational technology in support of all activities of the Center. There are four major areas of research focus in Core 1 of NCIBI: 1) Computational tools, 2) Data integration, 3) Software integration, and 4) Cross core collaboration.

Since inception, the Computation and Data Integration Core has made advances using native XML database technologies (TIMBER) to achieve deep integration and display of molecular interactions of proteins with provenance and uncertainty information (the Michigan Molecular Interactions Database; http://mimi.ncibi.org/). We have significantly improved our SAGA sub-graph matching algorithms for pathways analyses, which previously had been limited by NP-hard limits.

Core 2: Bioinformatics and Knowledge Acquisition

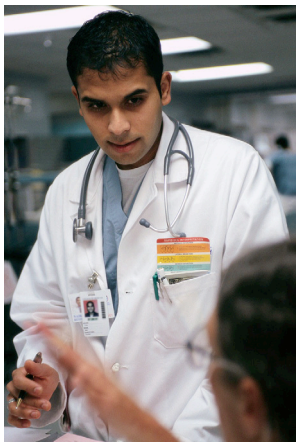
Modern molecular biology is increasingly driven by the need to analyze complex systems using modeling. As described in greater detail in the Core 3 DBPs, projects begin with an exploratory data analysis phase in which relevant information is identified and gathered from a wealth of resources in databases and text. In the second phase, an iterative process is pursued in which models are formulated, parameterized with established facts, evaluated to understand the implications of the model, and revised appropriately. During this modeling phase, additional data is typically gathered and parameters used in modeling are refined. Finally, a point is reached at which multiple models are consistent with all of the available data, and there's a need to perform new experiments to confirm one model and reject alternatives. The goal of Core 2, and ultimately this center, is to provide the information and software environment needed to accomplish this process rapidly and reliably.



The Bioinformatics Core (Core 2) brings together a diverse and multidisciplinary group of collaborators—inside, and outside of the University of Michigan. There are three major areas of activity for Core 2: 1) development of bioinformatics tools, 2) information retrieval and integration, 3) Software Integration and Cross-Core Collaboration.

Advances in the Bioinformatics and Knowledge Acquisition Core have emerged from development and application of Bayesian network tools, automated natural language processing (NLP), information retrieval capabilities (MiSearch, Clairlib), and large-scale exploration and data-mining of the biomedical literature (PubViz). In cooperation with our strategic partners, the NCIBI has utilized and built upon such well-supported platforms as GenePattern, Cytoscape, Protégé, and the Ensembl DAS infrastructure.

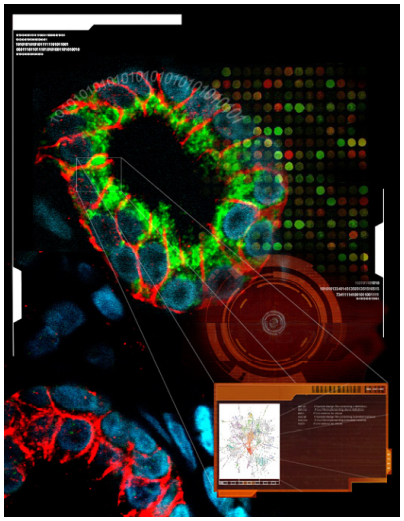
Core 3: Driving Biological Problems



The NCIBI driving biological problems are common diseases with complex and heterogeneous etiology. Complex diseases are the next frontier in biomedical research, requiring tools and workflows to aid researchers in understanding the rapidly growing scope of data and literature available to address their questions, including genetic, behavioral, and environmental variables. In order to mine and assimilate the wide range of emerging information, tools must be sophisticated, user-friendly, and generate insightful models of relevant disease processes. The DBPs provide insight into specific problems which can be addressed computationally and provide first-line review and feedback for testing the tools and applications developed by Cores 1 and 2; what we now call the NCIBI "Suite of Tools" (miBLAST, SAGA, MiMI, Oncomine, Bubble, SNP portal, NLP, and the imported GenePattern workflow and Sakai portal). We share our methodologic progress via live and archived webcasts of our weekly Tools and Technologies Seminars, as described in Core 6, as well as regular all-cores meetings, and project specific "Jamborees".

The tools developed by Cores 1 and 2, together with laboratory and community data resources, have accelerated our assembly of relevant information for research on our four Driving Biological Problems (DBPs): 1) for prostate cancer progression, we have a progressively annotated 976-gene androgen-regulated network, downstream targets for EZH2 polycomb proteins and have implemented an ambitious plan to study the newly-discovered TMPRSS2/ETS fusion genes; 2) on organ-specific complications of Type-1 Diabetes, we have extensive characterization of NF-kB involvement in glomerular injury and a remarkably improved platform for molecular research on neuropathy; 3) for heterogeneity of Type-2 Diabetes, our colleagues as part of the Fusion project, have played a leading role in the SNP and HapMap findings of several new disease-associated genes, and families are enrolled in genome-wide association studies of quantitative metabolic traits; and 4) a valuable collaborative resource has been created for studies of genetic susceptibility and phenotypic sub-classification of Bipolar Disorder.

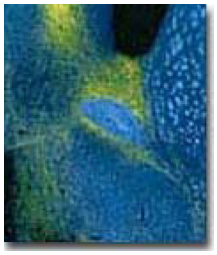
Progression of Prostate Cancer



Prostate cancer is a leading cause of cancer-related death in American men, second only to lung cancer. The major clinical judgments relate to distinguishing benign prostatic hypertrophy from prostate cancers and then recognizing and treating successive stages of intra-epithelial neoplasia, localized cancer, invasive cancer, and metastases. Most early stage cancers are androgen-dependent and therefore susceptible to treatments that block androgen action. As tumors progress, they tend to become androgen-independent as they develop invasive and metastatic properties. Current diagnostic efforts mostly utilize prostate-specific antigen (PSA) as a serum biomarker, tumor histopathology classified by Gleason score (scale of 10) for malignant attributes, and extent of tumor by various clinical methods. Several years ago the University of Michigan Complex Systems and Urology Programs organized a national workshop, including Drs. Chinnaiyan and Omenn, and published a paper on the complexity of prostate cancer biology with a focus on trying to predict the prognosis for patients with the highly heterogeneous Gleason 7 score prostate cancers. In the subsequent five years, advances in molecular biology and informatics have made it feasible to ask much deeper questions about these critical properties of prostate cancers.

Our overall hypothesis posits that global gene expression analysis will elucidate a select set of genes that dictate whether a clinically localized prostate cancer exhibits a progressive or non-progressive phenotype and that global protein studies (proteomics) will help identify tumor and circulating proteins as biomarkers. We recognize that the underlying predispositions from genetic and environmental/nutritional/hormonal mechanisms will be heterogeneous, so that multiple patterns of carcinogenesis and of response to treatments must be expected and differentiated. Some of these molecular patterns will be specific to prostate cancer, while others will be shared by subsets of cancers arising in other organs.

Complications of Type 1 Diabetes



Type 1 diabetes, also known as juvenile diabetes or insulin-dependent diabetes, is typically diagnosed in children and in young adults. Type 1 diabetes accounts for 3% of new cases of diabetes each year, and the cause is unknown. Complications from type 1 diabetes include cardiovascular disease, neuropathy, nephropathy, and retinopathy.

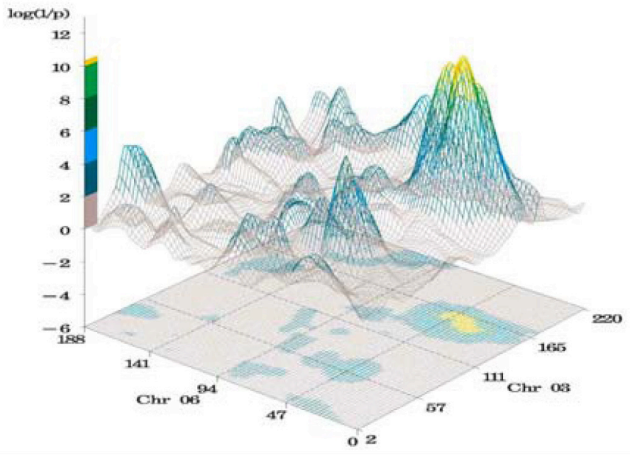
This DBP collaboration is establishing interactions between the University of Michigan Juvenile Diabetes Research Foundation (JDRF) Center and the National Center for Integrative Biomedical Informatics (NCIBI) in a long-term relationship that will be sustained through future funding, create new research initiatives in type 1 diabetes, and develop models for biological processes relevant to research on the complications of type 1 diabetes.

Population Genetics of Type 2 Diabetes

Type 2 diabetes (T2D) is a common, complex disease. While disease prevalence varies with age, gender, and population, it has been estimated that 6.6% of the U.S. population aged 20–74 suffers from T2D, also known as adult-onset or non-insulin-dependent diabetes mellitus (NIDDM). Similar rates of T2D have been observed in Finland and, in fact, T2D results in substantial morbidity and mortality worldwide. In the U.S., it has been estimated that diabetes is responsible for nearly 1/7 of all health care expenditures. There is overwhelming evidence for the familiality of T2D, and strong evidence for a genetic component. Genetic studies of T2D and T2D-related quantitative traits provide a promising approach to identify T2D susceptibility variants. Identification of these genetic variants should result in better understanding of disease etiology, and may help to identify at-risk individuals, suggest new drug targets and therapies, and assist in disease prevention. It also opens the way to a new level of inquiry in the genetic epidemiology of T2D, in which the joint effects of specific genetic variants and environmental factors can be examined more directly.

This new level of inquiry into the complex influences on T2D susceptibility requires the use of multiple heterogeneous datasets, mining of thousands of text references for patterns of information that may be relevant to T2D research, modeling of relationships found, and presentation of newly developed knowledge in a form that is easy for humans to grasp. Development of National Center for Integrative Biomedical Informatics (NCIBI) resources and integrating them with statistical genetics methods is improving the efficiency and completeness of work in T2D research and is providing direction and motivation to maximize the effectiveness of the NCIBI resources
for all researchers.

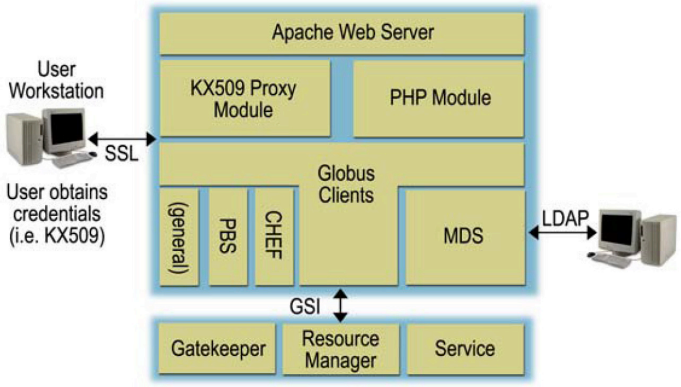
Genetics of Susceptibility to Bipolar Disorder



Bipolar Disorder (manic-depressive disorder) is characterized by severe and often debilitating swings of mood, ranging from mania (pathological elevations of mood with seemingly limitless and misguided energy) through depressive phases with low moods and no energy. The study of the genetics of bipolar disorder has become an expressed priority of the National Institute for Mental Health (NIMH). The National Center for Integrative Biomedical Informatics (NCIBI) resources are being used to advance the exploration of genetic and other mechanisms relating to susceptibility to bipolar disorder. Comprehensive integration of data and information from a wide range of sources is under way, and includes genetics, biochemistry, physiology, psychology, and phenomenology; searches are across species, and include orthologous genes and behaviors in these species. Experience using this Driving Biological Problem (DBP) is being used to enhance and train the capabilities of NCIBI.

Core 4: Infrastructure

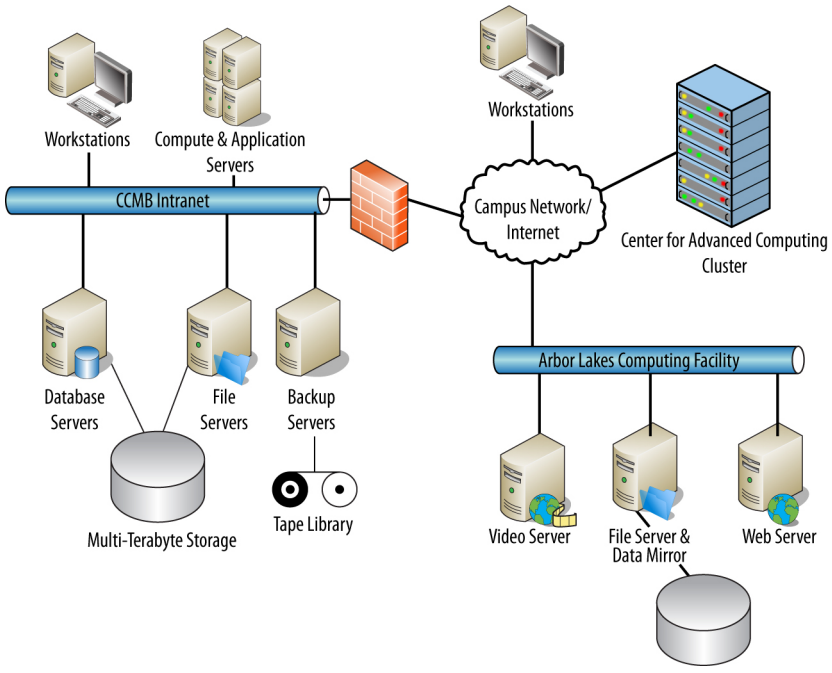
Core 4 is responsible for providing direct technical support to the other cores and creating a high-performance computing (HPC) and data infrastructure which links both local and national partners. This infrastructure serves as a means for developing, deploying, and using software tools and databases.



The Center recently acquired and installed a significant storage system as part of the NCIBI computing infrastructure, and was set up to work together with existing filer server appliance to create file shares that are available across the network to servers and workstations. The gateways allow us to create virtual file servers for user home directories and shared group project directories. The virtual file servers integrate with our Windows Domain to enforce authentication, authorization, and access control on files. The files are presented on the network using network protocols that make the files equally accessible from Windows desktops or Linux-based high performance computing servers. This capability increases user efficiency and reduces storage requirements.

In addition to storage, we improved our computing and web serving infrastructure. The Center acquired and installed additional computing and web server hardware. The infrastructure core aims to provide high-performance but cost-effective computing platforms for NCIBI. Thus, core 4 purchased multi-processor, multi-core computers that can be used for applications ranging from simple, single threaded tasks to parallel tasks. To make the most efficient use of this hardware, we have begun using server virtualization software. This software allows us to run multiple virtual machines on a single computer. Each VM can be configured to the needs of the specific application. We plan to extend this use of VMs to provide a centrally managed computing environment that meets the inherent variability of the various applications being developed and deployed by NCIBI.

In addition to computation, a primary goal of NCIBI is the integration of data from multiple sources. To support this integration, NCIBI has deployed database servers running Microsoft SQLServer and Oracle. We have been working with the various software and database developers in NCIBI to configure and tune these servers to best meet the requirements of the deployed applications; utilizing the skills of a full-time database administrator to provide professional database server administration and to work directly with the database developers to make the most efficient use of these resources.



Core 5: Education and Training

The NCIBI Education and Training Core leverages the placement of the NCIBI into the UM Center for Computational Medicine and Biology (CCMB), which also houses the Bioinformatics Graduate Program (BGP). This environment, in centrally appointed space at the University of Michigan with state-of-the-art video teleconferencing facilities, and has proved to be a magnet for NCIBI training opportunities for graduate students, post-doctoral fellows, and participating faculty across the campus and throughout the country.

Core 5 is 1) providing interdisciplinary training for all levels of NCIBI personnel; 2) establishing innovative education and training programs that are leveraging the Center's integrated information and computational resources, facilitating collaborative activities, and fostering community accessibility and data sharing with service to NCIBI developers, researchers, and users; 3) beginning to educate the NIH researcher and user community on use of the NCIBI systems and tools, including best practices in the representation of experimental data and in the construction of the knowledge access and data analysis environment; 4) educating the community about data sharing and software dissemination (also Core 6); and 5) assessing and evaluating educational and training programs (by Core 6).

NCIBI has successfully embarked on an innovative interdisciplinary training program for Graduate Student Research Assistants in bioinformatics, computer and information sciences, and biostatistics. NCIBI and NCBC tools are being actively integrated by NCIBI faculty into the pre-dissertation training curriculum of the students in these programs. This is serving as an ideal test-bed for broader dissemination (Core 6). We plan to mount an NCIBI post-doctoral program to complement these graduate training efforts, and we are actively recruiting Post Doctoral Fellows. These training opportunities will provide cross-training with a customized "mentorship team" matched to trainee needs. An NCIBI Certificate Program tailored to CS and IS concentrators has been approved and is being developed, which will leverage the UM Bioinformatics Graduate Program.

In collaboration with Core 6, NCIBI hosts a year-round weekly seminar series, the NCIBI "Tools and Technologies" seminars (Thursday, 12 Noon ET). Its focus is on innovative computational methods with DBP applications within NCIBI, and between NCIBI and its subcontractors and national outreach and Collaborating R01/R21 partners. Presentations are made by and for the working team, including students, post-docs, faculty and staff. These seminars are broadcast live by streaming video/audio over the Internet to a growing set of partners and interested users (in collaboration with Core 6).

Core 6: Outreach and Dissemination

This core is responsible for implementing and supporting the data sharing and software dissemination plans of National Center for Integrative Biomedical Informatics (NCIBI). One of these tasks is creating and managing the tools that NCIBI researchers need to share data and software both internally and externally. Core 6 is also responsible for executing the specific software dissemination and data sharing policies established by the Center. Core 6 is also tasked with outreach activities to the general scientific community. The core 6 staff have began gathering contact information for professional and scientific societies and groups, institutions, and others that may have an interest in the Center's work and results. As the system moves beyond its initial design and development stages and becomes ready for more wide usage, we will initiate contact with these groups to announce Center services. By actively seeking meetings and societies, we will also provide a list of potential presentation venues for Center investigators. Our planned dissemination activities should aid in the recruitment of new DBP co-investigators for the later years of the Center's activities.

There are four major areas of research focus in Core 6 of NCIBI: 1) NCIBI collaboration portal (web site); 2) cyber-infrastructure implementation for outreach, interaction and leveraging subcontractors and other partners; 3) collaborative R01, R21 and other grant programs; and 4) outreach highlights and plans.

NCIBI is disseminating these tools, data and means of integration, through portal-enhanced outreach and innovative web-based interactive training and educational programs for our partners around the country and for the broader NIH community and potential new collaborators. These lively education and training programs are available by streaming video at http://www.ncibi.org in easily downloadable form, with training module enhancements being added periodically. Use statistics indicate growing access by NIH collaborators and users nation-wide. NCIBI has been a leader in working with other NIH National Biomedical Computing Centers (NCBCs). In addition, NCIBI has collaboratively created and maintains the NCBC-wide portal ncbcs.org. NCIBI is also poised to leverage such other important NIH-funded programs as the Biomedical Informatics Research Network (BIRN), the cancer Biomedical Informatics Grid (caBIG), and the Clinical and Translational Sciences Award (CTSA) informatics networks.

Leadership and Contact Information for NCIBI

Brian D. Athey, Ph.D.
Principal Investigator
Senior Scientific Director
Core 6 & 7 Director, Core 5 Co-Director
Core 1 Investigator
3600 Green Court Suite 700
Ann Arbor, MI 48105-1570
Phone: 734-615-5914
Fax: 734-998-8571
Email: bleu@umich.edu

Hosagrahar V. Jagadish, Ph.D.
Senior Scientific Director
Core 1 Director
2238 EECS Building
Ann Arbor, MI 48109-2122
Phone: 734-763-4079
Fax: 734-763-8094
Email: jag@umich.edu

Gilbert S. Omenn, M.D., Ph.D.
Senior Scientific Director
Core 3 Director
A510C MSRB I
1150 W. Medical Center Dr.
Ann Arbor, MI 48109-0626
Phone: 734-763-7583
Fax: 734-647-8148
Email: gomenn@umich.edu

Barbara E. Mirel, Ph.D.
Core 5 Director
2112 School of Information North
1075 Beal Avenue
Ann Arbor, MI 48109-2112
Phone: 734-764-1858
Email: bmirel@umich.edu

James D. Cavalcoli, Ph.D.
NCIBI Project Manager
3600 Green Court Suite 700
Ann Arbor, MI 48105-1570
Phone: 734-615-5510
Fax: 734-998-8571
Email: cavalcol@umich.edu

Katrina Ward Hope
Director of Administration
Center for Computational Medicine and Biology
University of Michigan
2017 Palmer Commons 2218
Phone: 734-936-1807
Fax: 734-615-6553
Email: wardhope@umich.edu

National Center for Integrative Biomedical Informatics (NCIBI)
The University of Michigan
Palmer Commons Room 2017 Palmer Commons
100 Washtenaw Avenue Ann Arbor, MI 48109-2218
Phone: 734-615-5510 Fax: 734-615-6553
web site: http://www.ncibi.org/ email: ncibiinfo@umich.edu

National Institutes of Health
Grant #U54 DA021519