Scientific Annual Report 2012

Virginia Bioinformatics Institute
Washington St. (0477)
Blacksburg, VA 24061
T: 540.231.2100 | F: 540.231.2606
web: www.vbi.vt.edu

Thank you for taking an opportunity to look over our scientific annual report. This report contains summaries of the research, core services, and education & outreach activities ongoing within the institute. The theme of team science established by our founding director, Dr. Bruno Sobral, continues to underpin the operational philosophy of the institute.

Over the past several years, under the direction of Dr. Skip Garner, the institute has transitioned towards having a much greater emphasis on biomedical research and this aspect is clearly evident in the annual report. We are delighted about emerging partnerships and scientific synergies with the recently established Virginia Tech Carilion Research Institute, and we are committed to expansion of collaborations with our scientific colleagues on the Blacksburg campus. Under the direction of our recently appointed scientific director, Dr. Christopher Barrett, we will continue to work aggressively to build on existing strengths and past accomplishments.

We now live in a scientific world that is informatics-centric and transdisciplinary in nature. Indeed, there are no longer well-defined distinctions among discovery domains within the biological, biomedical, engineering and social sciences. Furthermore, the cost of data analysis now often exceeds the cost of data acquisition. From this perspective, future growth within VBI will necessarily involve an increased capacity for the analysis of complex interacting systems that is unconstrained by any disciplinary boundary. We are excited about past accomplishments and future prospects for the continued scientific growth of the institute, as well as VBI's contributions to outreach, education and core services. Finally, speaking for all the scientific investigators, we are extremely proud and appreciative of our extraordinary support staff whose contributions to the institute are truly significant.

Director, Life Science
Director, Virginia Bioinformatics Institute


ACIL
ADVANCED COMPUTING + INFORMATICS LABORATORY

SUMMARY

The Advanced Computing and Informatics Laboratories (ACIL) is an organization of team-science laboratories in VBI focused on advancing informatics methods in massively interacting systems. The programmatic theme connecting the topical foci of the separate lab enterprises is employment of high-performance, pervasive, and data intensive computing-based approaches to complex biological and socially coupled systems. The affiliated laboratories in the ACIL include the Nutritional Immunology and Molecular Medicine Laboratory (NIMML), Network Dynamics and Simulation Science Laboratory (NDSSL), and the Social and Decision Informatics Laboratory (SDIL). Domain research topics range over, and often integrate, systems biology to detailed regional population behavior and economic modeling. In addition, ACIL labs engage in related basic research in mathematics, theoretical computer science, network science and network-centric high-performance computing systems.

The Nutritional Immunology and Molecular Medicine Laboratory (NIMML) conducts translational research aimed at developing novel therapeutic and prophylactic approaches for modulating immune and inflammatory responses. The NIMML is composed of over 20 researchers and combines computational modeling, bioinformatics approaches, pre-clinical experimentation and human clinical trials to elucidate the mechanisms of immune regulation at mucosal surfaces. This knowledge will ultimately accelerate the development of novel treatments for infectious and immune-mediated diseases.

The Network Dynamics and Simulation Science Laboratory (NDSSL) integrates analytical and computational research across diverse domains in an overarching program. The aim is to develop synthetic information systems and associated analytical methods appropriate for very large complex systems. In addition to basic research and informatics technology development, translational efforts involving the commercial sector to put the synthetic methods into wide use are also a part of the NDSSL portfolio. Sponsors currently include NSF, NIH, DoD, IARPA, and others. The synthetic information technology program references the following research domain topics: systems biology for enteric immunology, social epidemiology of contagious diseases, contagious disease public health policy and surveillance, public health economics, large detailed population modeling, human social cognition and social behavior, disaggregate demand-side economic modeling and their various detailed markets, societal infrastructure modeling, policy decision support, crowdsourced surveillance, and crowdsourced population science.

The SDIL is a policy informatics-enabled, analysis-oriented laboratory. Its current focus is on DoD-related and public health policy domain topics. The SDIL helps produce valuable informatic system requirements for decision and system analysis problems they confront. Reciprocally, as technology and information is available from the basic technology research in other ACIL laboratories, it is incorporated in SDIL analysis projects.

THE NUTRITIONAL IMMUNOLOGY AND MOLECULAR MEDICINE LABORATORY

The NIMML has over the last five years received over $12 million in externally funded programs, including NIH RO1-level and Center level funding as well as corporate support, with about $2 million of research expenditures accrued and 20 papers published during the reporting period. The NIMML team leads the NIH-funded Center for Modeling Immunity to Enteric Pathogens (MIEP, www.modelingimmunity.org), a $10.6 million program directed by Dr. Josep Bassaganya-Riera with the goal of modeling the regulatory mechanisms underlying immune responses to gastroenteric pathogens. The MIEP engages 30 scientists and is part of a $45 million infectious disease program (Modeling Immunity for Biodefense). The long-term goal of the MIEP program is to characterize the mechanisms of immunoregulation underlying mucosal immune responses to gastroenteric pathogens. More

specifically, we have characterized new host pathways controlling gut inflammation during infection with *Helicobacter pylori,* enteroaggregative *Escherichia coli* and *Clostridium difficile.* We integrated computational and experimental approaches to investigate the role of these gut bacteria as harmless or beneficial commensals versus pathogens.

Mechanisms of Immune Modulation by Abscisic Acid

Under the grant entitled "Mechanisms of Immune Modulation by Abscisic Acid" awarded by the NIH/NCCAM (Grant No R01AT004308) to Dr. Raquel Hontecillas and Dr. Bassaganya-Riera, the NIMML investigated the mechanisms of immune modulation by abscisic acid (ABA) and discovered an alternative mechanism of activation of peroxisome proliferator-activated receptor gamma (PPAR γ) and anti-inflammatory activity by ABA. This and other discoveries resulted in 20 published papers, 10 patents filed or issued and 10 invited presentations.

In previous work, the NIMML team demonstrated that ABA, a naturally occurring compound derived from plants, has beneficial effects on several conditions and diseases including obesity-related inflammation, diabetes, atherosclerosis, and inflammatory bowel disease. One idea for how ABA reduces inflammation in these instances is that it binds to a special region of PPAR γ, a binding site known as the ligand-binding domain where the drug would be expected to latch on to and exert its effect. Our results show that this is not the case and, for the first time, we have demonstrated that ABA works independently of this ligand-binding domain of the receptor. The outcomes of this research illustrate the synergism that can result from combining computational and experimental approaches to characterize therapeutic targets. By using molecular modeling approaches, we were able to identify potential binding sites for ABA on the lanthionine synthetase C-like 2 (LANCL2), a protein required for the beneficial health effects of ABA **(Fig. 1)**. We have also shown that ABA affects the expression of several genes involved in inflammation, metabolism and cell signaling, which provides further clues for possible intervention points in the treatment of inflammatory, infectious and immune-mediated diseases.

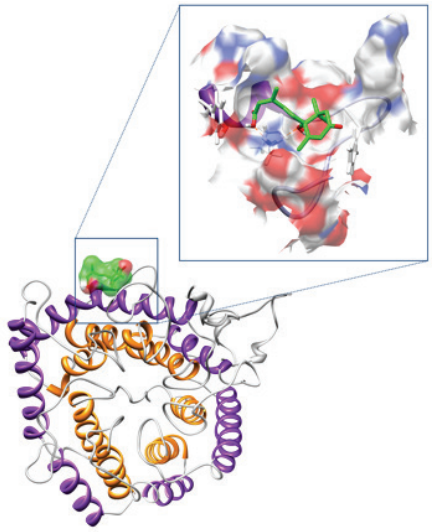


Figure 1. Energetically favorable pose for abscisic acid (ABA) bound to lanthionine synthetase C-like 2 (LANCL2) in ribbon representation with helix layers colored in purple (outer) and orange (inner). ABA in green stick representation surrounded by molecular surface with coloring by element. Insert: D278 to N296 of LANCL2 in ribbon and surface representations, colored by element, with residues Y279, K284, and F285 displayed

In a follow-up to our previous research on ABA, we investigated its ability to modulate host responses and health outcomes during influenza viral infection. Influenza accounts for anywhere from 3,000 to 49,000 deaths per year in the United States alone, according to the Centers for Disease Control. It is difficult to treat if not caught immediately; antivirals usually become ineffective after the incubation period has passed, and resistance to antiviral drugs poses a serious public health problem in the face of outbreaks. ABA was shown to be most effective at about seven to ten days into the infection, targeting the immune response rather than the virus itself, which is a safer way to reduce flu-associated fatalities. Most drugs for respiratory infections target the virus rather than the inflammatory responses caused by the virus. ABA suppresses lung inflammation and damage through activation of LANCL2, a newly identified pathway. This research led by Dr. Hontecillas was published in the Journal of Nutritional Biochemistry.

Crohn's Disease Clinical Trial on the Efficacy of Conjugated Linoleic Acid

The NIMML first reported the anti-inflammatory efficacy of conjugated linoleic acid (CLA) in a pig model of colitis in 2002. We next characterized the mechanism of action of CLA *in vivo* in mice. Under a contract from Cognis (now BASF) awarded to the NIMML, we discovered important new information on the efficacy of CLA in treating human Crohn's disease (CD), a form of inflammatory bowel disease (IBD). CLA is a naturally occurring fatty acid found in meat and dairy products known for its anti-cancer and immune modulatory properties. In collaboration with the Division of Gastroenterology and Hepatology at University of North Carolina School of Medicine and the Wake Forest Medical Center, the NIMML researchers found that CD patients who took supplementary CLA showed noticeable improvement. CLA was administered as a supplement in study subjects with mild to moderate CD for 12 weeks. We found a marked improvement

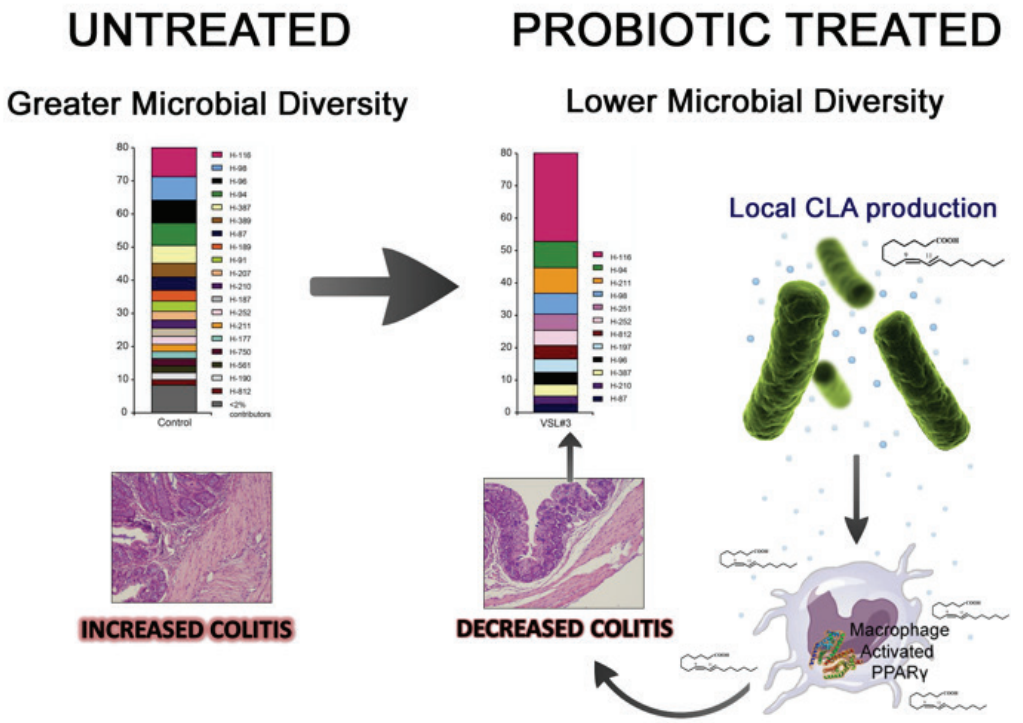


Figure 2. Proposed model for a mechanism of action underlying the protective effects of probiotic bacteria in mouse models of gut inflammation and cancer. Colonization with probiotic bacteria modulates gut microbial diversity and favors local production of conjugated linoleic acid (CLA) in the colon that targets myeloid cell peroxisome proliferator-activated receptor γ (PPAR γ) to suppress colitis.

in disease activity and quality of life, and no adverse side effects since CLA was well tolerated by all of the study subjects. CLA significantly suppressed the ability of peripheral blood CD4+ and CD8+ T cell subsets to produce IFN-γ, TNF-α and IL-17 and to proliferate at week 12. In summary, CLA represents a promising new supportive intervention for gut inflammation. This is in contrast with the results of human clinical studies using n-3 polyunsaturated fatty acids in IBD that remain largely unimpressive. The present study has shed new light on the clinical potential of this compound and provided insights on the possible mechanisms of immune modulation targeted by CLA in the human system. Based on these results, a larger Phase II double-blind, placebo-controlled, randomized trial with several doses of CLA is warranted.

In a related project, we have demonstrated that gut probiotic bacteria produce CLA locally in the gut where it contributes to modulating the composition of gut bacteria and decreasing infiltration of inflammatory cells such as macrophages in the intestinal wall of mice with IBD and inflammation-associated colorectal cancer. Thus, changes in microbial diversity and local CLA production are implicated in PPAR γ-dependent mechanisms of action underlying the anti-inflammatory and anti-carcinogenic effects of probiotic bacteria **(Fig. 2)**. This novel mechanistic model is supported by: results of loss-of-function analyses illustrating the requirement of macrophage PPAR γ in mediating the full spectrum of anti-inflammatory effects of probiotic bacteria in the gut; *in vivo* evidence indicating a reduction of colonic bacterial diversity with a marked predominance of TRF H-116 and local CLA production in colons of probiotic-treated mice; and remarkable similarities in the ability of probiotic bacteria and CLA to modulate macrophage function at the gut mucosa.

CENTER FOR MODELING IMMUNITY TO ENTERIC PATHOGENS

Beneficial effects of *Helicobacter pylori* in obesity and diabetes

Helicobacter pylori is a Gram-negative, microaerophilic bacterium of the Epsilonproteobacteria that colonizes the stomach as a dominant member of the gastric microbiota of nearly half of the world's population. *H. pylori*'s phylogenetic tree reflects the major human migration out of Africa, across Europe, through Asia, and into the New World. Gastric infection with *H. pylori* has been associated with various gastric diseases, including gastritis, peptic ulcer disease, gastric adenocarcinoma, and gastric mucosa-associated lymphoma.

While *H. pylori* infection is associated with severe disease, mounting evidence suggests a beneficial role of this bug in chronic inflammatory, allergic or autoimmune diseases. Specifically, there also is increasing evidence of *H. pylori* protection against esophageal and cardial pathologies, childhood asthma, childhood allergies, obesity and diabetes. The mechanisms underlying this protective effect of *H. pylori* acting as a commensal bacterium or a pathogen are largely unknown.

Interestingly, there is an inverse secular trend between the incidence of obesity and gastric colonization with *H. pylori*, a bacterium that can affect the secretion of gastric hormones that relate to energy homeostasis. Previously, the NIMML team characterized the whole genome and function of *H. pylori* V225d, cultured from a Venezuelan Piaroa Amerindian, an atypical strain found in areas with low incidence of *H. pylori*-associated cancer. During this reporting period, we demonstrated that colonization with *H. pylori* exerts beneficial effects in obesity and diabetes. Specifically, we examined metabolic and inflammatory markers in genetically obese/diabetic (db/db) mice and mice with diet-induced obesity experimentally infected with isogenic forms of *H. pylori* strain 26695: the *cag PAI* wild type and its *cag PAI* mutant strain 99-305. *H. pylori* colonization decreased fasting blood glucose levels, increased levels of leptin, improved glucose tolerance and suppressed weight gain. A response found in both wild type and mutant *H. pylori* strain-infected mice included decreased white adipose

tissue macrophages (ATM) and increased adipose tissue regulatory T cells (Treg) cells. Gene expression analyses demonstrated upregulation of gastric PPAR γ-responsive genes (i.e., CD36 and FABP4) in *H. pylori*-infected mice. The loss of PPAR γ in immune and epithelial cells in mice impaired the ability of *H. pylori* to favorably modulate glucose homeostasis and ATM infiltration during high fat feeding. Thus, gastric colonization with some commensal strains of *H. pylori* ameliorates glucose homeostasis in mice through a PPAR γ-dependent mechanism and modulates macrophage and Treg cell infiltration into the abdominal white adipose tissue.

These data suggest that colonization by *H. pylori* strains lacking the *cag PAI* could provide partial protection against some metabolic disorders. Thus, if this theory holds true, the disappearance of *H. pylori* in developed countries may be a contributing factor to the epidemics of obesity and diabetes. Future studies will examine the mechanisms by which specific *H. pylori* strains modulate regulatory and effector pathways in the gastric mucosa, and their correlation with improvements of chronic inflammatory diseases. Ongoing efforts under the MIEP program involve creating mathematical and computational models of the effector and regulatory pathways triggered by *H. pylori* at the gastric mucosa. The better understanding of such pathways will shed new light on the dual role of *H. pylori* as a pathogen and a commensal organism.

Modeling immune responses toward *Clostridium difficile*

Clostridium difficile typically is a harmless environmental sporulated Gram-positive anaerobic bacterium, but it has recently re-emerged as a significant enteric pathogen implicated in nosocomial diarrhea, colitis and even death, particularly after antibiotic treatment. *C. difficile* grows in the intestine of individuals with altered commensal microflora due to treatment with antimicrobials, immunosuppressants, cytostatic agents or proton pump inhibitors. An increase in both incidence and severity of *C. difficile*-associated disease (CDAD) has been reported over the last few years.

Previously, CDAD was a concern in older or severely ill patients, but the emergence of new hypervirulent strains such as NAP1/BI/027 has resulted in increased morbidity and mortality for other age groups in the United States, Canada and Europe. The increased virulence of *C. difficile* is attributed to greater sporulation and production of binary toxins or to higher level of fluoroquinolone resistance. Persistent or severe CDAD is currently being treated with discontinuation of the antibiotic therapy that led to the disease and vancomycin therapy. Nevertheless, these therapeutic approaches do not restore the normal microflora and are not effective in clostridial clearance, but further prolong *C. difficile* shedding and destroy beneficial gut anaerobic bacteria. In contrast to targeting the bacterium and its toxins directly, a better understanding of the cellular and molecular basis underlying the host response will enable the rational development of host-targeted therapeutics for CDAD.

We have investigated the mechanisms underlying PPARγ modulation of mucosal immune responses to *C. difficile*, including a possible relationship between nuclear receptors and micro RNAs. Specifically, we applied mathematical and computational modeling approaches in combination with mouse challenge studies to elucidate the mechanisms underlying the interactions between PPARγ activity and miRNA-146b and how this interaction regulates colitis during *C. difficile* infection **(Fig. 3)**. Also, we have investigated how genetic or pharmacological manipulation of PPARγ modulates colonic inflammatory cytokines and effector T helper (Th17) responses to *C. difficile* infection in mice. We have found that T cell PPARγ prevents colitis and down-modulates effector T cell responses in mice with CDAD, suggesting a potential crosstalk between miRNAs and the PPAR γ pathway.

In silico simulations of miRNA and PPARγ in host responses to *C. difficile* infection in the colon

To further characterize the interactions occurring between *C. difficile*, miR-146b and PPARγ, we developed a computational and mathematical model of the colonic gene expression changes occurring in the colon following *C difficile* infection.

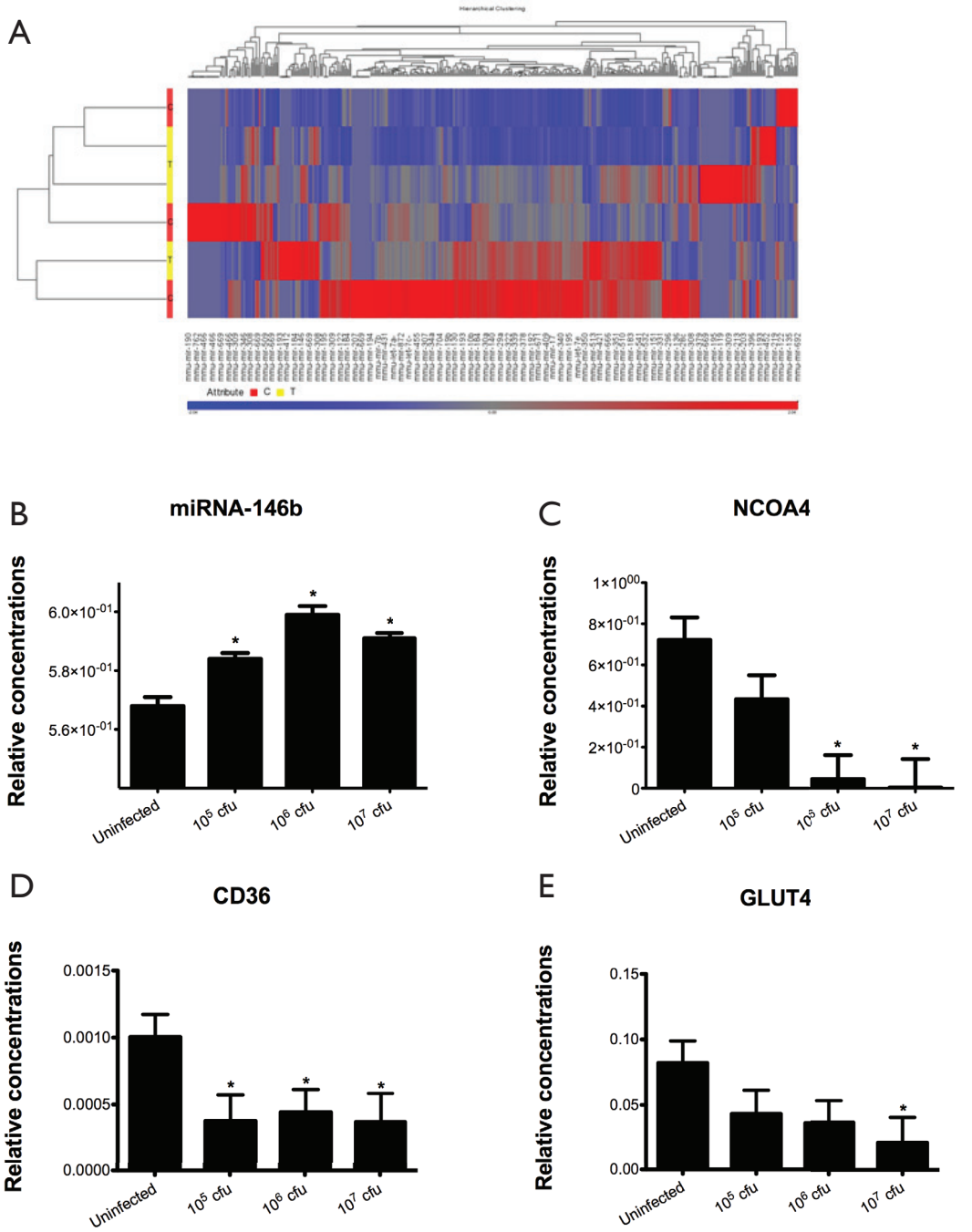


Figure 3. Effect of *Clostridium difficile* infection on the colonic expression of micro RNA-146b and target genes NCOA4, CD36 and GLUT4 mRNA in mice. MiRNA-seq heatmap of *Clostridium difficile*-infected (T) and uninfected control (C) mice (A). Colonic expression of miRNA-146b (B) as well as NCOA4 (C), CD36 (D) and GLUT4 (E) were assessed by real-time quantitative RT-PCR in mice infected with *C. difficile* (n=10). Data are represented as mean ± standard error. Points with an asterisk are significantly different when compared to the control group ($P<0.05$).

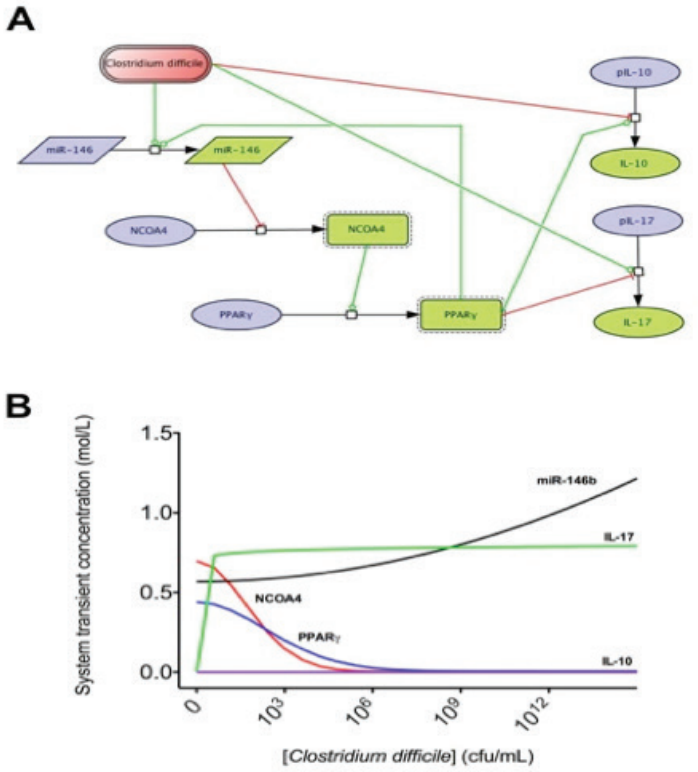


Figure 4. Computational modeling of mucosal immune responses to *Clostridium difficile* infection. (A). CellDesigner-based illustration of the COPASI model of the mucosal immune response to *Clostridium difficile* mucosal immune response. The model represents the interaction between *C. difficile*, miRNA-146, nuclear receptor coactivator 4 (NCOA4), peroxisome proliferator-activated receptor γ (PPAR γ), interleukin 10 (IL-10) and interleukin 17 (IL-17) in Systems Biology Markup Language format. Inhibition is represented in red and activation in green. (B) COPASI steady state scan showing the variation on the species concentrations with increasing computational concentration of *C. difficile*. *In silico* simulations show how increasing concentrations of *C. difficile* increase miRNA-146b levels, thus decreasing NCOA4 and PPAR γ. In line with the experimental data, IL-17 expression also increases with the infection.

This network was constructed based on our experimental findings and literature information. By using this model, we explored the mechanisms by which *C. difficile* modulates the expression of effector and inflammatory cytokines. Our computational simulation predicted an upregulation of miR-146b, and IL-17 and a down-regulation of NCOA4 and PPARγ in colons of *C. difficile*-infected mice **(Fig. 4).**

We have used loss-of-function approaches in combination with pharmacological activation of PPARγ and computational modeling to investigate for the first time the critical role of PPARγ in regulating immune responses and disease severity following *C. difficile* infection. Our data suggest that overexpression of miRNA-146b in the colon might exacerbate inflammatory responses by suppressing PPARγ activity through a mechanism involving suppression of NCOA4, a co-activator molecule required for activation of PPAR γ. These results suggest novel host-targeted drug development strategies for CDAD.

ENISI framework

ENteric Immunity SImulator (ENISI) is an agent-based model designed to simulate the cell movement and interactions in the gastroenteric mucosa. Here we discuss the current 1.5 version of ENISI. In ENISI each individual cell is modeled and acts independently.

ENISI modeling environment

First, we developed ENISI, a modeling system for inflammatory and regulatory immune pathways initiated by microbe-immune cell interactions in the gut. ENISI is an interaction-based model where individual cells are modeled, along with their movement though different tissues, and the probabilistic outcomes of cell-cell interaction. ENISI has the ability to simulate at least 107 individual cells. With ENISI, mucosal immunologists will be able to test and generate hypotheses for enteric disease pathology and propose interventions through experimental infection of an *in silico* gut. This is done by using a simple scripting language to assign parameter values that conform to knowledge and assumptions of the experimental scenario to be simulated. Simulation outcomes given different experimental conditions allow observation of *in silico* behaviors that are not readily seen through *in vitro* and *in vivo* techniques. This information can then be used to generate novel treatment strategies that can be tested in the laboratory. The ENISI modeling environment has already been illustrated by developing an *in silico* model and dynamic simulation of *H. pylori* and a simulation of dysentery resulting from *Brachyspira hyodysenteriae* infection so as to identify aspects of the host immune pathways that lead to continued inflammation-induced tissue damage even after pathogen elimination. The ENISI modeling environment has been designed from the start for scalability, efficiency and the use of large-scale parallel computing systems. The underlying message-passing middleware is based on the EpiSimdemics modeling environment. Our recent work has demonstrated scaling to between 10^7-10^8 cells using 600-800 cores of a parallel cluster. To our knowledge, this is the first individual agent-based model of an immune system that achieves this kind of scalability.

We have studied the performance issues of ENISI and developed several improvements at the algorithmic and programming levels. We describe the improved results in light of the previous version (v1), which was published during the IPDPS 2012 conference. Also, we improved many of the C++ functions, streamlining the code to provide programmatic improvements. In the following figure, we show the improvements in terms of speed, execution time and scaling. As the model is large enough to be run on a single cluster, we compared the speed with a 12-core simulation. Here we tested ENISI for 107 cells on up to 700 processing nodes. We have seen a remarkable increase in speed which has allowed us to make more runs of the simulation than the previous version. We can now do multiple runs of each simulation, which will facilitate sensitivity analysis.

We integrated the ENISI Tool into our website (www.modelingimmunity.org), which allows the scientific community to study the behavior of *H. pylori* infection in a mouse stomach. Users are able to request an account and set up their own *in silico* experiments, which can be submitted to VBI's high-performance computing cluster. The user is informed by email of job completions and can view the results on the MIEP website. We integrated the statistical results calculated by multiple simulations of the ENISI *H. pylori* model into the MIEP website, which are accessible via the simulation results under the ENISI tool (http://www.modelingimmunity.org/modeling/enisi/).

ENISI Visual

There are two versions of ENISI: ENISI high-performance computing (HPC), described above, and ENISI Visual. ENISI Visual is implemented using the REPAST platform and can be used to develop and calibrate ENISI models at a smaller scale before running them in our HPC cluster.

ENISI Visual is adapted from the HPC implementations of ENISI and emphasizes quality user interfaces and visualizations **(Fig. 5)**. ENISI Visual allows users to specify the initial concentrations of cells, the pathogen infectious dose and simulation time. From the interface, users can control the animation speed and specify the output charts, figures, real-time animations, snapshots, and videos. ENISI Visual has integrated secretion and propagation of cytokines and chemokines, as well as cell movement models. The real-time animations during *in silico* experiments allow users to quickly test hypotheses and discover novel phenomena. ENISI Visual will be enhanced to include host-pathogen-

microbiota interactions. ENISI Visual provides rich graphic user interfaces. Users can control initial cell concentrations, simulation speed, data and graphic outputs. A video of ENISI Visual is available at http://www.modelingimmunity.org/modeling-tools/enisi-visual/

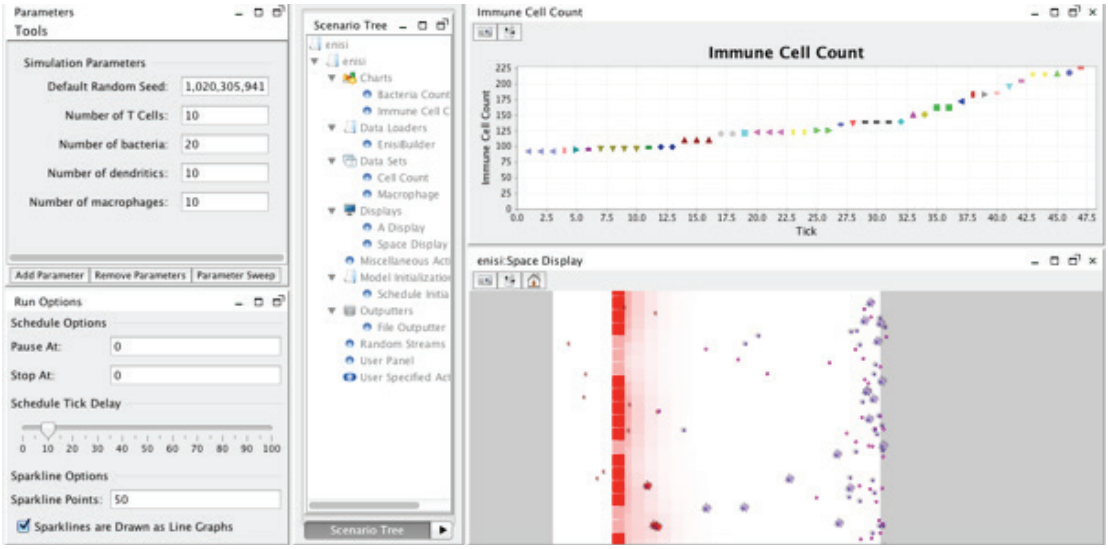


Figure 5. ENISI Visual User Interface. Users can specify initial cell and pathogen concentrations, simulation speed, output charts, figures and videos.

Student and post-doctoral training

The NIMML is training seven PhD students in the Genetics, Bioinformatics and Computational Biology (GBCB) program. The GBCB students in the NIMML receive training at the interface of immunology, bioinformatics and computational modeling and simulation of biological processes. The NIMML has over the last five years trained 10 PhD students, 40 undergraduate students and nine post-doctoral associates.

During this period, the NIMML students reported the following achievements:

- Nikki Lewis, one of the PhD students whom Dr. Bassaganya-Riera is co-advising with Dr. David Bevan in biochemistry, received a NIH F31 award.
- Adri Carbo, a second year PhD student in GBCB, received a first poster award in the MIB immune modeling Symposium at the University of Rochester School of Medicine for her modeling studies related to CD4+ T cell differentiation.
- Monica Viladomiu, a first-year PhD student in GBCB, received a travel award from the American Association of Immunologists to present her *Clostridium difficile* research.
- Each NIMML PhD student has co-authored at least two papers and up to five papers during this reporting period.

The NIMML has been selected by the American Association of Immunologists (AAI) as a component of their AAI high School Teachers Summer Research Program. This program provides immunology learning and research opportunities to a select group of high school teachers in the U.S. In 2012, Dr. Bassaganya-Riera mentored Mr. Stephen Biscotte, a teacher from Cave Spring High School in Roanoke.

Business development and commercialization efforts

The NIMML research on anti-inflammatory product development resulted in the creation of a company in 2008. BioTherapeutics, Inc., develops new therapies for chronic inflammatory diseases

from nature's innovations. BTI has established a management team, including an experienced CEO, Erika Smith, and a scientific advisory board to advance anti-inflammatory drug development efforts towards commercialization.

Future directions

The NIMML has consolidated as a major research laboratory at VBI within the ACIL Division and with fully integrated computational and experimental capabilities. The MIEP program and other planned centers and programs will position Virginia Tech at the forefront of computational immunology, -omics and immunoinformatics research and discovery. The NIMML will play a major role in this Virginia Tech-wide transdisciplinary effort.

The NIMML in collaboration with a team of investigators at Virginia Tech will submit a program P01 proposal to the National Center for Complementary and Alternative Medicine (NCCAM) to establish the Center of Excellence in Natural Anti-inflammatories (CENA). This new Center will create new systems-level knowledge of the mechanisms of action underlying the preventative effects of natural products in gastrointestinal diseases. The proposal will be submitted in September 2013. The CENA is highly synergistic with the Virginia Tech Center for Drug Discovery (VTCDD) led by University Distinguished Professor Dr. David Kingston, for which Dr. Bassaganya-Riera a founding member.

Publications

Bassaganya-Riera J, Viladomiu M, Pedragosa M, de Simone C, Carbo A, Shaykhutdinov R, Jobin C, Arthur JC, Corl B, Vogel H, Storr M, and Hontecillas R. Probiotic bacteria produce conjugated linoleic acid locally in the gut that targets macrophage PPAR γ to suppress colitis. *PLOS ONE*. 2012; 7(2):e31238.

Bassaganya-Riera J, Hontecillas R, Horne WT, Sandridge M, Herfarth H, Bloomfeld R, and Isaacs K. Conjugated linoleic acid modulates immune responses in patients with mild to moderately active Crohn's disease. *Clinical Nutrition*. 2012; 31(5):721-727.

Bassaganya-Riera J, Viladomiu M, Pedragosa M, de Simone C, and Hontecillas R. Immunoregulatory mechanisms underlying prevention of colitis-associated colorectal cancer by probiotic bacteria. *PLOS ONE*. 2012; 7(4):e34676

Bassaganya-Riera J, Dominguez-Bello MG, Kronsteiner B, Carbo A, Viladomiu M, Pedragosa M, Lu P, Zhang X, Sobral BW, Mane S, Mohapatra SK, Horne WT, Guri AG, Groeschl M, Lopez-Velasco G and Hontecillas R. *Helicobacter pylori* infection activates gastric PPAR γ and ameliorates glucose homeostasis in mice. *PLOS ONE*. In press. 2012.

Bassaganya-Riera J, Hontecillas R, and Isaacs KL. Implementation of integrative approaches to improve the design of Crohn's disease clinical trials. *Clinical Nutrition*. In press. 2012.

Bisset K, Bassaganya-Riera J, Carbo A, Eubank S, Hontecillas R, Hoops S, Marathe M, Mei Y, Wendelsdorf K, Xie D, Yeom J, Alam M. High-performance interaction-based simulation of gut immunopathologies with ENISI. In IPDPS 12. Shanghai, China. 2012.

Coursodon-Boyiddle C, Snarrenberg C, Adkins-Rieck C, Bassagnya-Riera J, Hontecillas R, Lawrence R, Brenna J, Jouni Z, and Dvorak B. Pomegranate seed oil reduces damage in a rat model of necrotizing enterocolitis. *American Journal of Physiology – Gastrointestinal and Liver Physiology*. 2012; 303(6): G744-751.

Hontecillas R, Roberts PC, Carbo A, Horne WT, Velayudhan B, and Bassaganya-Riera J. Abscisic acid ameliorates influenza virus-related pulmonary pathology and down-regulated MCP-1 through a PPAR γ-dependent mechanism. *Journal of Nutritional Biochemistry*. In press. 2012.

Hontecillas R and Bassaganya-Riera J. Expression of PPAR gamma in intestinal epithelial cell is dispensable for the prevention of colitis by dietary abscisic acid. *e-SPEN Journal*. In press. 2012.

Lu P, Hontecillas R, Horne WT, Carbo A, Viladomiu M, Pedragosa M, Bevan DR, Lewis SN, and Bassaganya-Riera J. Computational modeling-based discovery of novel classes of anti-inflammatory drugs that target lanthionine synthetase component C-like 2. *PLOS ONE*. 2012; 7(4): e34643.

Mei Y, Hontecillas R, Zhang X, Bisset K, Eubank S, Hoops S, Marathe M, and Bassaganya-Riera J. (2012) ENISI Visual, an agent-based simulator for modeling gut immunity. In: *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM 2012). Philadelphia, PA.

Viladomiu M, Hontecillas R and Bassaganya-Riera J. Nutritional protective mechanisms against gut inflammation. *Journal of Nutritional Biochemistry*. In press. 2012.

Viladomiu M, Hontecillas R, Hoops S, Pedragosa M, Michalak P, Michalak K, Guerrant RL, Roche JK, Warren CA and Bassaganya-Riera J. Modeling immune responses to *Clostridium difficile* by peroxisome proliferator-activated receptor γ and miRNA-146b. *PLOS ONE*. In press. 2012.

Wendelsdorf KW, Alam M, Bassaganya-Riera J, Bisset K, Eubank S, Hontecillas R and Marathe M. ENteric Immunity SImulator: A tool for *in silico* study of gastroenteric infections. *IEEE Trans. NanoBioScience*. 2012; 11(3):273-288.

Presentations

Bassaganya-Riera J. Modeling immunity to enteric pathogens. Rochester Summer School and Symposium. Rochester, NY. 2012.

Bassaganya-Riera J. Predictive modeling of CD4+ T cell responses during enteric pathogen infection. Mount Sinai/Yale School of Medicine Immunology Seminar. New York, NY. 2012.

Bisset K, Bassaganya-Riera J, Carbo A, Eubank S, Hontecillas-Magarzo R, Hoops S, Marathe M, Mei Y, Wendelsdorf K, Xie D, Yeom J, Alam M. High-performance interaction-based simulation of gut immunopathologies with ENISI. 26th IEEE International Parallel & Distributed Processing Symposium. Shanghai, China. 2012.

Carbo A, Hontecillas R, Hoops S, Kronsteiner B, Lu P, Wendelsdorf K, Mei Y, Eubank S, Marathe M, and Bassaganya-Riera J. PPAR γ activation drives Th17 cells into a Treg phenotype. American Association of Immunologists Annual Meeting. Boston MA. 2012.

Kronsteiner B, Bassaganya-Riera J, Washington C, Vento S, Pedragosa M, Viladomiu M, Carbo A and Hontecillas R. *Helicobacter pylori* infection in pigs is dominated by a Th1 and cytotoxic immune response. American Association of Immunologists Annual Meeting. Boston, MA. 2012.

Lu P, Hontecillas R, Viladomiu M, Pedragosa M, Carbo A, Bevan DR, Lewis SN, and Bassaganya-Riera J. Immunomodulatory actions of lanthionine synthetase C-like 2-based drugs, American Association of Immunologists Annual Meeting. Boston, MA. 2012.

Mei Y, Hontecillas R, Zhang X, Bisset K, Eubank S, Hoops S, Marathe M, and Bassaganya-Riera J. ENISI Visual, an agent-based simulator for modeling gut immunity. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2012). Philadelphia, PA. 2012.

Viladomiu M, Hontecillas R, Pedragosa M, Michalak P, Michalak K, Guerrant RL, Roche JK, Warren CA and Bassaganya-Riera J. Modulation of immune responses to *Clostridium difficile* by peroxisome proliferator-activated receptor γ and miRNA-146b. American Association of Immunologists Annual Meeting. Boston, MA. 2012.

Washington C, Bassaganya-Riera J, Viladomiu M, Pedragosa M, Guerrant RL, Roche JK and Hontecillas R. Th17 responses driven via PPAR γ blockade lead to faster recovery from enteroaggregative *Escherichia coli* infection. American Association of Immunologists Annual Meeting. Boston, MA. 2012.

NETWORK DYNAMICS AND SIMULATION SCIENCE LABORATORY

The Network Dynamics and Simulation Science Laboratory (NDSSL) has made substantial progress in basic science and the development of usable tools to study large complex systems. Over the last eight years, NDSSL has received grants and contracts worth over $70 million. Thirty-three million dollars of funded programs are currently in progress. Extramural expenditures over the last fiscal year were to the tune of nine million dollars. The work was funded by a diverse set of organizations, including: the National Institutes of Health, Department of Transportation (through a subcontract from AECOM), the Centers for Disease Control, the Department of Defense, the Bill and Melinda Gates Foundation (through a subcontract from Swiss Tropical and Public Health Institute), the National Science Foundation, the Department of Homeland Security, The Department of Energy, the Lawrence Livermore National Laboratory, and the Department of Interior as a grant via the Intelligence Advanced Research Projects Activity (IARPA) and VT internal funds. NDSSL has published over 200 articles in peer-reviewed journals, conferences and books. These articles have appeared in top venues including Nature, PNAS, Infocom, Sigmetrics, JACM, Mobihoc, AAMAS, IEEE TON, AAAI, SC, ICS and WSC. Seven students have completed their PhDs so far and we expect 15 more students to complete their doctoral dissertation in the next three years. The group is comprised of 50+ members that include students, postdoctoral fellows, staff scientists and faculty members. Currently, NDSSL employs a total of 40 students and postdoctoral fellows that include: 26 PhD students, nine masters students, two undergraduates and three postdoctoral fellows. Additionally, the laboratory has employed 50 students and five postdoctoral fellows over the last eight years. The students have spanned various departments across the university, including the Departments of Computer Science, Economics, Electrical and Computer Engineering, Industrial and Systems Engineering, Statistics, Mathematics, Population Health Science, Physics and the GBCB program.

We have strengthened our presence in the National Capital Region and play a leading role in a new institutional initiative in Policy Informatics for Complex Systems. We are pursuing new programs in wireless networks, commodity markets, computational economics, energy systems, sustainable interdependent infrastructure design and analysis, and high-performance computing. We continue to develop advanced high-performance computing-based computational tools and methods for reasoning about complex systems. These resources are integrated into web services providing synthetic databases, national scale interaction-based simulations, and analysis tools. We have used these tools in several stakeholder-designed studies supporting policy planning for epidemics and other large-scale national and international incidents. We continue to build upon and enhance our simulation development processes and tools that were initially developed for the Defense Threat Reduction Agency (DTRA) of the Department of Defense, called the Comprehensive National Incident Management System (CNIMS). CNIMS integrates surveillance, simulation-assisted hypothesis testing, and decision support for use in situational awareness and planning in complex systems. We continue to successfully transition our technology so that we can rapidly expand upon our platform to simulate new events based on defined event criteria. The novel use of high-performance computing assets for addressing practical societal problems represents a unique capability developed by NDSSL. This is coupled with the ability to process vast quantities of data from surveys, as well as online social media and interactive tools to support data and theory driven decision support.

Scientific Progress

The Network Dynamics and Simulation Science Laboratory (NDSSL) continues to make significant scientific progress in the fields of simulation and network science, high-performance computing, public health epidemiology, wireless networks and theory. Here, we highlight some of the achievements in the reporting period.

- We have developed a high-performance data intensive modeling environment to study co-evolving socially-coupled systems. A key feature of the modeling and analysis environment is its ability to represent and study cascading effects between physical societal infrastructures as well as the ability to represent adaptive human behaviors. The modeling environment has been used to study the impact of a large-scale human-initiated crisis. The modeling environment is based on the idea of synthetic information systems and uses a network-intensive computing paradigm for achieving the desired scalability.
- Using the aforementioned modeling and decision support environment, we have carried out the first detailed computational experiment to study human behaviors and interactions between people and societal infrastructures in the aftermath of a large explosion caused by an improvised nuclear device. While several studies have been published on this topic in the past, very little work has been done on understanding how individuals, families and organizations are likely to react and the composite effect of this adaptive multi-scale behavior. We studied health outcomes and infrastructure usage as individuals panic, try to reconstitute their families, evacuate and find medical help. Our results clearly demonstrate the role of communication networks, specifically wireless communications and social media in improving the overall outcome. The results suggest that an improved and resilient wireless network can provide users location-aware information and allow them to better coordinate with their families. Government services can also reach individuals and provide important and timely information that reduces the overall panic in society, improves the ability to reconstitute families and find medical help. A second important outcome of the study is the identification of modeling epochs. Specifically, our results show that behavioral, policy and infrastructure representations to study the impact of such an overwhelming event must be consistent with the evolving situation. Furthermore, the modeling environment should identify when a certain class of representations is appropriate and adequate and when they are no longer needed.
- **Synthetic data:** We have continued to refine the synthetic population of 250 million individuals endowed with daily activity patterns performed at specific, real locations. New methods were developed for refining activities for subpopulations such as tourist and business travelers to an area. Additionally, mapping and locating buildings, roads and electrical and cellular infrastructure have been enhanced to provide a more robust real world environment. We have begun developing international synthetic populations and networks. Initial models are now available for New Delhi, Sweden and Israel. These models will be refined. New models are also being developed for the UK. These synthetic data sets played a crucial role in our studies for the Department of Defense and the National Institutes of Health.
- **Scalable middleware:** We continued the development of Simfrastructure – a software middleware that supports resource management, digital libraries and interfaces and high-performance models. Simfrastructure forms the basis of the high-performance computing modeling and simulation environment that is the core of CNIMS.
- **Modeling environments:** We have developed new modeling environments. Epicure is a high-performance computing-based modeling environment to study the spread of malware in wireless networks. It has been shown to scale to urban regions with more than three million devices. Epicure was used to study various

control measures. Intersim is a modeling environment to study social contagion. Like Epicure, Intersim is a high-performance computing-based modeling environment and has been used to develop methods for blocking and enhancing contagions in large social networks. Another modeling environment has been developed for understanding the spatio-temporal spectrum demand. Using this tool, we have studied market-based mechanisms for spectrum sharing. The work was cited in a recent PCAST report on spectrum sharing. We have also developed an interactive modeling environment called Indemics to support infectious disease epidemiology. Indemics is the first high-performance modeling environment that supports interactivity; users can start and stop model execution and interrogate the simulation state.

- **Algorithmics:** We have obtained the first sequential and distributed algorithms to compute various service metrics in wireless networks under a realistic model of radio interference. These algorithms have provable guarantees in terms of time and space. They also have provable worst-case performance guarantees. We have developed new algorithms and heuristics to study control and optimization problems arising in social network contagions, e.g. fads, ideologies, etc.
- **Social media data driven modeling:** We have begun methods for synthesizing various social networks based on social media. These networks have then been used to understand critical social media users. The ideas are being used to forecast social and political events in Latin America.
- **Forecasting and situation assessment:** We have developed model-based techniques for forecasting epidemic characteristics such as the time to peak, the total infection rate and number of infections at the peak.

We are extending our modeling frameworks and tools to address other problems in systems biology and medical and health informatics. We have made significant progress in developing high-performance computing-based computing for highly resolved models for systems biology. For example, we created the ENteric Immunity Simulator (ENISI), a simulator of the gastrointestinal (GI) tract mucosal immune responses created for generating and testing hypothesis of mechanisms of immune regulation in response to the presence of resident commensal or pathogenic bacteria. We are also extending our methods in social networks and modeling diffusion processes for developing models of individual behaviors within a networked environment as they pertain to non-infectious health issues such as smoking and obesity.

Programmatic Progress

The work of the NDSSL has resulted in progress in several key program areas. We received a new National Science Foundation (NSF) grant to develop computational foundations of socio-technical network science and engineering. Anil Vullikanti, one of our team members, received a Department of Energy Early Career award to study contagions in social networks. We also received a NSF grant to study efficient market-based frequency allocation mechanisms in wireless radio networks. We have received a basic research and development grant from the DTRA for a project to develop mathematical and computational foundations pertaining to verification and validation of large-scale agent-based models for socio-technical systems. Recently, we also received a new NSF grant that aims to study game theoretic aspects in computational epidemiology. Another NSF contract was received at the beginning of the reporting period to develop a scalable and extensible cyber-infrastructure to support network science.

We continue to work on the DTRA-funded project on modeling the spread of infectious diseases related to civil and military planning for situation assessment and response analysis. The contract has been extended for five additional years. We have received excellent reviews from our customers. We continue to contribute as a principal institution in the NIH Modeling Infectious Disease Agent Study (MIDAS) project. This project was also renewed as a part of a competitive process for five more years.

National and International Leadership

The NDSSL team members participate on various program committees and panels including, but not limited to: the Yale New Haven Center for Emergency Preparedness and Disaster Response Study to Determine the Requirements of a National Operational Epidemiological Modeling Process; the Modeling Subcommittee of the Biomedical Research Workforce working group for the Advisory Committee to the Director (NIH); and the Population Modeling Working Group of the NIH's Multiscale Modeling Consortium. We received recognition "for exceptional service and dedication in providing technical expertise and guidance to the Director, NIH" for our participation in the National Institutes of Health Blue Ribbon Panel for Risk Assessment of the National Emerging Infectious Disease Laboratory at Boston University Medical Center; the Review Committee for Computing Environment and Life Sciences (CELS), University of Chicago; Argonne National Laboratory LLC, the Scientific Advisory Board; SMART Infrastructure Facility, University of Wollongong, Australia; the Scientific Review Committee, Global System Dynamics Coordination; Future and Emerging Technologies Programme, European Commission; and the external advisory committee for the Institute of Scientific Exchange in Torino, Italy. A member of the NDSSL team also served on the Language Foundation Selection Committee for the Lagrange Prize.

We also continue to advise DTRA and the Department of Defense on modeling and policy planning related to interdependent infrastructure analysis and pandemic preparedness.

In the reporting period, we have also hosted a training workshop for the DTRA and personnel from operational military commands on the DIDACTIC tool developed by our team.

Review panels: National Science Foundation, French National Science Foundation, National Institutes of Health, and MIT-Skoltech Foundation

Program committee members: The 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), The 9th International Conference on the Evolution of Language (Evolang 2012), ACM Compute 2012; The 31st IEEE Conference on Computer Communications (INFOCOM), 2012; The 8th IEEE International Conference on eScience (eScience 2012); Military Communications Conference 2012; The 5th International Conference on Simulation Tools and Techniques (SIMUTools), 2012; The International Conference on Parallel Processing (ICPP), 2012 ;1st International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), 2012; The 7th International Symposium on Algorithms for Sensor Systems, Wireless Ad Hoc Networks and Autonomous Mobile Entities (Algosensors).

Editorial boards: Data Science, Guest editor of Journal of Computer Networks.

Plenary and invited talks

1. Facing the Challenge of Infectious Diseases: Integrating mathematical modeling, computational thinking, and ICT applications. Courmayeur, Italy. 2012.
2. Institute for Systems Science and Health. St. Louis, MO. 2012.
3. American Public Health Association's annual meeting, organized session "Models to Support Public Health Practice: MIDAS." San Francisco, CA. 2012.
4. Invited talk. Tepper School of Business, Carnegie Mellon University. Pittsburgh, PA. 2012.
5. ASSET 2012. Invited talk honoring Prof. Fernando Vega-Redondo. Cyprus. 2012.
6. International Conference on Networks in Biology, Social Science and Engineering, Indian Institute of Sciences. Madras, India. 2012.

Workshop co-organization: We co-organized a number of interdisciplinary workshops on topics related to policy informatics, socio-technical network science and agent-based modeling including:

- AAAI Fall Symposium, Social Networks and Social Contagion. Washington, DC. October 2012. http://staff.vbi.vt.edu/swarup/snsc/
- Current Challenges in Computing 2012: Network Science, Institute for Scientific Computing Research (ISCR) at Lawrence Livermore National Laboratory (LLNL). http://www.regonline.com/builder/site/default.aspx?EventID=1091457.

Education and outreach activities

NDSSL had significant achievements in Education and Outreach in the reporting period. This included activities for students ranging from high school all the way to doctoral students. Members of our group also taught courses for the Department of Mathematics and the Department of Computer Science at Virginia Tech during the spring and fall of 2012.

- We co-organized the Flu-Tracker exhibit at the USA Science and Engineering Festival 2012 that expanded on the exhibit from 2010. The exhibit included displays on the entire public health policy planning cycle, ranging from mathematical modeling to drug discovery.
- We co-organized a special year series titled Mathematics of Networks from August 2011 to July 2012 (http://math.iisc.ernet.in/imi/Networks.php). A number of workshops and tutorials were organized as part of this series, bringing together researchers from diverse backgrounds. We developed two short courses that were given at the Indian Institute of Sciences and the Lawrence Livermore National Laboratory.

Completed PhD dissertations

- Jonathan P. Leidig, Computer Science. "Epidemiology experimentation and simulation management through scientific digital libraries." He is currently an Assistant Professor at Grand Valley State University, MI.
- Elaine O. Nsoesie. Computational Biology and Bioinformatics. "Sensitivity analysis and forecasting in network epidemiology models." She is now a postdoctoral fellow at Virginia Tech and Harvard Children's Hospital.
- Katherine Wendelsdorf. Computational Biology and Bioinformatics. "Models of the mucosal inflammatory and regulatory immune pathways: The role of host response in microbial persistence and pathogenesis." She is now a postdoctoral fellow at NIH.
- Guanhong Pei, Electrical and Computer Engineering. "Distributed scheduling and delay-throughput optimization in wireless networks under the physical interference model." He will join Qualcomm.
- Karthik Channakeshava, Electrical and Computer Engineering. "High-performance scalable and expressive modeling environment to study mobile malware in large dynamic networks." He is now at Ericsson.

A paper written by one of our students, Gaurav Tuli, received the best student paper award at the 2012 International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. A poster presented by our student Zhao Zhao received the best poster award at the PhD Forum at the 2012 IEEE 26th International Parallel and Distributed Processing Symposium. Another poster by group members with our students won the best poster award at Informs 2012.

Publications

Alexander K, Lewis B, Marathe M, Eubank S, and Blackburn J. Modeling of wildlife-associated zoonoses: Applications and caveats. *Vector Borne Zoonotic Dis.* 2012; 12(12), 1005-1018.

Barrett C, Channakeshava K, Huang F, Kim J, Marathe A, Marathe M, …Vullikanti A. Human initiated cascading failures in societal infrastructures. *PLOS ONE*. 2012; 7(10), e45406.

Bisset K, Aji A, Marathe M, and Feng W. High-performance biocomputing for simulating the spread of contagion over large contact networks. *Bmc Genomics*. 13. 2012.

DuBois T, Eubank S, and Srinivasan A. The effect of random edge removal on network degree sequence. *Electronic Journal of Combinatorics*. 2012; 19(1).

Han B, Hui P, Vullikanti A, Marathe MV, Shao JH, and Srinivasan A. Mobile data offloading through opportunistic communications and social participation. *IEEE Transactions on Mobile Computing*. 2012; 11(5), 821-834.

Istrate G, Marathe M, and Ravi S. Adversarial scheduling in discrete models of social dynamics. *Mathematical Structures in Computer Science*. 2012; 22(5), 788-815.

Khan M, Kuhn F, Malkhi D, Pandurangan G, and Talwar K. Efficient distributed approximation algorithms via probabilistic tree embeddings. *Distributed Computing*. 2012; 25(3), 189-205.

Lee B, Yilmaz S, Wong K, Bartsch S, Eubank S, Song Y, . . . Huang S. Modeling the regional spread and control of vancomycin-resistant enterococci (VRE). *Clinical Microbiology and Infection*. In press. 2012.

Marathe A, Pan Z, and Apolloni A. Analysis of friendship network and its role in explaining obesity. *ACM Transactions in Science and Technology, Special Issue on Social Computing*. 2012.

Nsoesie E, Beckman R, and Marathe M. Sensitivity Analysis of an Individual-Based Model for Simulation of Influenza Epidemics. *PLOS ONE*. 2012; 7(10), e45414.

Wendelsdorf K, Alam M, Bassaganya-Riera J, Bisset K, Eubank S, Hontecillas R, . . . Marathe M. ENteric Immunity SImulator: A tool for *in silico* study of gastroenteric infections. *IEEE Transactions on Nanobioscience*, 2012; 11(3), 273-288.

Refereed conferences

Abdelhamid S, Alo R, Arifuzzaman S, Beckman P, Bhuiyan M, Bisset K, . . . Zhao Z. CINET: A cyberInfrastructure for network science. Presented at the 8th IEEE International Conference on eScience. Chicago, IL. October 8-12, 2012.

Aji A, Dinan J, Buntinas D, Balaji P, Feng W, Bisset K, and Thakur R. MPI-ACC: An integrated and extensible approach to data movement in accelerator-based systems. Presented at the IEEE International Conference on High-performance Computing and Communications (HPCC). Liverpool, UK. June 25-27, 2012.

Aji A, Panwar L, Ji F, Chabbi M, Murthy K, Balaji P, . . . Thakur R. On the efficacy of GPU-integrated MPI for scientific applications. Presented at the 2012 IEEE 26th International Parallel and Distributed Processing Symposium. Shanghai, China. May 21-25, 2012.

Barrett C, Beckman R, Bisset K, Chen J, Eubank S, Vullikanti A, . . . Srinivasan,A. Optimizing epidemic protection for socially essential workers. Presented at the ACM International Health Informatics Symposium. IHI 2012, Miami, FL. January 28-30, 2012.

Bisset K, Aji A, Bohm E, Kale L, Kamal T, Marathe M and Yeom J. Simulating the spread of infectious disease over large realistic social networks using Charm++. Presented at the 17th International Workshop On High-Level Parallel Programming Models and Supportive Environments (HIPS). Shanghai, China. May 21, 2012.

Bisset K, Alam M, Bassaganya-Riera J, Carbo A, Eubank S, Hontecillas R, . . . Marathe M. High-performance interaction-based simulation of gut immunopathologies with ENteric Immunity Simulator (ENISI). Presented at the 2012 IEEE 26th International Parallel and Distributed Processing Symposium. Shanghai, China. May 21-25, 2012.

Bisset K, Eubank S, and Marathe M. High-performance informatics for pandemic preparedness. Paper presented at the 2012 Winter Simulation Conference. Berlin, Germany. December 9-12, 2012.

Deodhar S, Bisset K, Chen J, Ma Y, and Marathe M. Enhancing user productivity and capability through integration of distinct software in epidemiological systems. Presented at the 2nd ACM SIGHIT International Health Informatics Symposium, Miami, FL. January 28-30, 2012.

Khan M, Pandurangan G, Pei G, and Vullikanti A. Brief announcement: A fast distributed approximation

algorithm for minimum spanning trees in the SINR model. Presented at the International Symposium on Distributed Computing (DISC), Salvador, Bahia, Brazil. October 16-18, 2012.

Kim J, Marathe A, Pei G, Saha S, Subbiah B, and Vullikanti A. Analysis of policy instruments for enhanced competition in spectrum auction. Paper presented at the IEEE Dynamic Spectrum Access Networks (DySPAN). Bellevue, Washington. October 16-19, 2012.

Kuhlman C, Vullikanti A, and Ravi S. *Controlling opinion bias in online social networks*. Paper presented at the WebSci '12 Proceedings of the 3rd Annual ACM Web Science Conference. Evanston, IL. June 22-24, 2012.

Lakkaraju K, Swarup S, and Gasser L. Consensus under constraints: modeling the great English vowel shift. Presented at the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP). University of Maryland, College Park, MD. April 3-5, 2012.

Liao S, Ma Y, Chen J, and Marathe A. Paid sick leave: Is it a good way to control epidemics? Presented at the 2nd International Conference on Complex Sciences: Theory and Applications. Santa Fe, NM. December 5-7, 2012.

Mei Y, Hontecillas R, Zhang X, Bisset K, Eubank S, Hoops S, . . . Bassaganya-Riera J. ENISI Visual, an agent-based simulator for modeling gut immunity. Presented at the IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia, PA. October 4-7, 2012.

Mortveit H. Limit cycle structure for block-sequential threshold systems. Presented at the Cellular Automata for Research and Industry (ACRI-ACA). Santorini, Greece. September 24-27, 2012.

Mortveit H and Macauley M. Cycle equivalence of finite dynamical systems containing symmetries. Presented at the Automata 2012. Corsica, France. September 19-21, 2012.

Pei G and Vullikanti A. Low-complexity scheduling for wireless networks. Paper presented at the 13th ACM international symposium on Mobile Ad Hoc Networking and Computing. Hilton Head, SC. June 11-14, 2012.

Pei G and Vullikanti, A. Distributed approximation algorithms for maximum link scheduling and local broadcasting in the physical interference model. Presented at the IEEE International Conference on Computer Communications (INFOCOM). Orlando, FL. March 25-30, 2012.

Pei G and Vullikanti, A. Brief announcement: Distributed maximum link scheduling under the physical interference model. Presented at the International Symposium on Distributed Computing (DISC). Salvador, Bahia, Brazil. October 16-18, 2012.

Saha S, Marathe A, Pei G, Subbiah B, Kim J, and Vullikanti A. Clearing secondary spectrum market with spatio-temporal partitioning. Paper presented at the IEEE Dynamic Spectrum Access Networks (DySPAN). Bellevue, Washington. October 16-19, 2012.

Swarup S and McCarthy C. Representational momentum may explain aspects of vowel shifts. Paper presented at the Artificial Life XIII. Michigan State University, East Lansing, MI. July 19-22, 2012.

Tuli G, Marathe M, Ravi S, Swarup S, and Gasser L. Addiction dynamics may explain the slow decline of smoking prevalence. Presented at the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP). University of Maryland, College Park, MD. April 3-5, 2012.

Youssef M and Scoglio C. Mitigation of epidemics in social networks through optimal contact adaptation. Presented at the International Conference on Network Science. Evanston, IL. June 20-22, 2012.

Zhao Z, Wang G, Butt A, Khan M, Kumar VS, and Marathe M. *SAHad: Subgraph analysis in massive networks using hadoop*. Presented at the 2012 IEEE 26th International Parallel and Distributed Processing Symposium. Shanghai, China. May 21-25, 2012.

Conferences and Posters based on abstracts:

Korkmaz G, Kuhlman C, Marathe A, Marathe M, and Vega-Redondo F. The role of communication via online social networks in the dynamics of collective action. Sunbelt XXXII (32) Social Networks Conference. Redondo Beach, CA. March 12-18, 2012.

Kuhlman C, Vullikanti A, Marathe M, Mortveit H, Ravi S, and Rosenkrantz D. Dependence of diffusion dynamics on network construction methods. Sunbelt XXXII (32) Social Networks Conference. Redondo Beach, CA. March 12-18, 2012.

Lewis B. Dynamic policy evaluation environment for disaster preparedness. Dynamics of Preparedness. Pittsburgh, PA. October 22-24, 2012.

Marathe M, Marathe A, Kuhlman C, Korkmaz G, and Vega-Redondo F. Emergence of collective action induced by social media common knowledge. Association of Southern European Economic Theorists 2012 Annual Meeting. Limassol, Cyprus. November 1-2, 2012.

Rivers C, Parikh N, Swarup S, Lewis B, and Eubank S. Behavior-driven simulations of disasters to foster resilient communities. American Public Health Association Annual Meeting. San Francisco, CA. October 27-31, 2012.

Youssef M, Parikh N, Swarup S, and Eubank S. Contact patterns between resident and transient populations in social networks. MIDAS Network Meeting. Arlington, VA. November 14-16, 2012.

SOCIAL DECISION INFORMATICS LABORATORY (SDIL)

The Social Decision Informatics Laboratory (SDIL) develops and uses analytical technology in areas such as public health, national and international security, and social and public policy. It provides a key mechanism for transitioning technologies from all labs under the ACIL umbrella into practical, actionable knowledge and technology for policy-makers.

SDIL works closely with the NDSSL to develop and use agent-based models of social networks to explain observed phenomena, provide insight into a range of future phenomena, and guide the course of action within the context of specific problems. Based on detailed simulation-based modeling and integration techniques, these models can provide government decision-makers with synthetic data sets that replicate how social and technological systems are affected, not only by physical laws, but also by human behavior, regulatory agencies, courts, government agencies, and private enterprises. Such models aid decision-makers in grappling with the vast amounts of data that often arise from interdependent, complex systems. Examples of collective decisions influencing networks on a massive scale can be seen by the regime changes that took place during the Arab Spring. The models and capabilities provided by SDIL can be invaluable tools that permit social and political decision-makers to better understand and explain, explore and take action to influence the millions of interacting agents that make up the network in their particular sector.

DIVISION CONTRIBUTORS

Abhijin Adiga, Ashwin Aji, Sudarshan Aji, Md. Maksudul Alam, S. M. Arifuzzaman, Chris Barrett, Richard Beckman, Josep Bassaganya-Riera, Md Hasanuzzaman Bhuiyan, David Bisaillon, Keith Bisset, Shridhar Chandan, Jiangzhuo Chen, Hannah Chung, Arron Dawson, Suruchi Deodhar, Sherief El Meligy, Stephen Eubank, Annette Feng, Jesun Firoz, Prasad Ganesh, Kevin Hall, Shamimul Hasan, Raquel Hontecillas, Jose Jimenez, Anurodh Joshi, Tariq Kamal, Rushi Kaw, Md-Abdul M Khan, Yasamin Khorramzadeh, Junwhan Kim, Gizem Korkmaz, Siddharth Krishnan, Barbara Kronsteiner-Dobramysl, Christopher Kuhlman, Edwina Lamm, Kathleen Laskowski, Bruce Lawlor, Jonathan Leidig, Bryan Lewis, Shaojuan Liao, Kristian Lum, Yifei Ma, Hemanth Makkapati, Akshay Maloo, Achla Marathe, Madhav Marathe, William Marmagas, Yongguo Mei, Gaurav Mishra, Henning Mortveit, Kalyani Nagaraj, Elaine Nsoesie, Ann Paul, Nidhi Parikh, Mireia Pedragosa Marin, Guanhong Pei, Thierry Platini, Caitlin Rivers, Daniel Rosenkrantz, Sudip Saha, Zalia Shams, Sara Shashaani, Sushrut Shirole, Paula Stretz, Balaaji Suna Pana Subbiah, Rajesh Subbiah, Samarth Swarup, Gaurav Tuli, Anil Vullikanti, Sandra Wagener, Katherine Wendelsdorf, Jody Williams, Mary Williams, Huadong Xia, Dawen Xie, Jae-seung Yeom, Ming Yi, Mina Youssef, Zhao Zhao


BIO
BIOSYSTEMS

SUMMARY

Although there are a variety of different opinions about the exact definition of "systems biology," all of them are unified by a focus on complex interactions that contribute to the functional operation of biological pathways, living cells, organs, organisms and populations. The Biosystems Division at VBI is comprised of an eclectic group of investigators having expertise ranging from computational modeling, statistical analysis, microbial gene expression and metabolism and the molecular basis for infection and disease as it relates to the host-pathogen-environment triangle. Many of the extramurally funded projects within the Biosystems group are integrated with other research programs within VBI or through active and lively collaborations with other groups at Virginia Tech or other research institutions.

DENNIS DEAN LABORATORY

Research in the laboratory of Dr. Dennis Dean is focused on two principal themes: the mechanism for biological nitrogen fixation and the biological pathways for assembly of simple and complex metalloclusters. With respect to nitrogen fixation, our group developed a combined biochemical-genetic approach to identify where substrates interact with nitrogenase, the biological catalyst of nitrogen fixation. This work enabled a series of biophysical approaches leading to the development of a comprehensive model for how certain substrates interact with the nitrogenase active site.

Very recently our laboratory, in collaboration with Dr. Lance Seefeldt at Utah State University, used a genetic approach to remodel nitrogenase such that it has the capacity to reduce carbon dioxide to yield methane and various short-chain high-value olefins. This finding suggests that the nitrogenase mechanism could provide clues for the rational development of metal-based catalysts for carbon dioxide sequestration.

The second theme has involved investigations on the biological assembly of iron-sulfur clusters. These simple inorganic structures are perhaps nature's most ancient prosthetic group and are likely to have contributed to the emergence of life on earth. Our laboratory discovered the biological mechanism for the assembly of iron-sulfur clusters and introduced the concept that both simple and complex iron-sulfur-containing prosthetic groups are pre-assembled on molecular scaffolds. Both *in vitro* and *in vivo* approaches have been used to establish the validity of the scaffold hypothesis. Both the mechanism of sulfur trafficking and the role of scaffolds in metallocluster assembly have now been established as nearly universal biological processes. Recent progress in this area has involved a collaborative project with Dr. Juan Fontecilla's group in Grenoble to elucidate the crystallographic structure of an iron-sulfur cluster biosynthetic complex captured in the process of cluster assembly. Ongoing work involves a comprehensive whole-genome transcriptome analysis of bacterial strains that are defective in various aspects of iron-sulfur assembly with the goal of understanding how this process is integrated with and contributes to intermediary metabolism.

Publications

George SJ, Barney BM, Mitra D, Igarashi RY, Guo Y, Dean DR, Cramer SP, Seefeldt LC. EXAFS and NRVS reveal a conformational distortion of the FeMo-cofactor in the MoFe nitrogenase propargyl alcohol complex. *J Inorg Biochem*. 2012; 112: 85-92. http://www.ncbi.nlm.nih.gov/pubmed/22564272

Liu Y, Dos Santos PC, Zhu X, Orlando R, Dean DR, Söll D, Yuan J. Catalytic mechanism of Sep-tRNA: Cys-tRNA synthase: sulfur transfer is mediated by disulfide and persulfide. *J Biol Chem*. 2012; 287: 5426-5433. http://www.ncbi.nlm.nih.gov/pubmed/22167197

Lukonyano D, Yang ZY, Barney BM, Dean DR, Seefeldt LC, Hoffman BM. Unification of reaction pathway and kinetic scheme for N2 reduction catalyzed by nitrogenase. *Proc Natl Acad Sci USA*. 2012; 109: 5583-5587.

Marinoni EN, de Oliveira JS, Nicolet Y, Raulfs EC, Amara P, Dean DR, Fontecilla-Camps JC. (IscS-IscU)2 complex structures provide insights into Fe2S2 biogenesis and transfer. *Angew Chem.* 2012; 51:5439-42. http://www.ncbi.nlm.nih.gov/pubmed/22511353

Mayweather D, Danyal K, Dean DR, Seefeldt LC, Hoffman BM. Temperature invariance of the nitrogenase electron transfer mechanism. *Biochemistry.* 2012; 51: 8391-8398. http://www.ncbi.nlm.nih.gov/pubmed/23050654

Seefeldt LC, Hoffman BM, Dean DR. Electron transfer in nitrogenase catalysis. *Curr Opin Chem Biol.* 2012; 16:19-25. http://www.ncbi.nlm.nih.gov/pubmed/22397885

Yang ZY, Moure VR, Dean DR, Seefeldt LC. Carbon dioxide reduction to methane and coupling with acetylene to form propylene catalyzed by remodeled nitrogenase. *Proc Natl Acad Sci USA.* 2012; 109: 19644-19648.

Yamanaka Y, Zeppieri L, Nicolet Y, Marinoni EN, de Oliveira JS, Odaka M, Dean DR, Fontecilla-Camps JC. Crystal structure and functional studies of an unusual l-cysteine desulfurase from *Archaeoglobus fulgidus*. *Dalton Trans.* 2012 Nov 19.

Invited seminars:

North Dakota State University (Department of Chemistry)

Rutgers University (Department of Microbiology)

Rice University (Department of Biochemistry

Conference lectures

Keynote lecturer. International conference on iron-sulfur proteins. Cambridge, England.

Keynote lecturer. Gordon Research Conference on iron-sulfur enzymes. Mount Holyoke, MA.

International conference on the role of metal hydrides in biology. Oxford, England.

Professional recognition

Fellow of the American Association for the Advancement of Science.

Fellow of the American Academy of Microbiology.

Purdue University College of Science Distinguished Alumnus.

Funded projects

Nitrogenase Mechanism.
National Institutes of Health
R01-GM59087

Maturation of Fe-S proteins in bacteria.
National Science Foundation
MCB-0717710

Broadening opportunities for non-traditional graduate students in biomolecular science.
National Science Foundation
STEM 085198

DICKERMAN LABORATORY

Dr. Dickerman's research continues to be focused in several key areas. These include a partnership with Techulon to develop peptide nucleic acid antimicrobial therapies, understanding protein interactions in wood development, and ongoing work in insertion mutagenesis in plants. As of November, Dr. Dickerman also works part time in the Data Analysis Core to provide analysis for large data sets.

Techulon

One important effort over 2012 was the contribution to a successful DARPA proposal to develop peptide nucleic acid antimicrobial therapies with Techulon. Personnel on the Techulon side were Joshua Bryson and Bret Malone. Dr. Dickerman's effort was devoted to predicting target genes for PNA knock-down by mining literature on essential genes across various bacterial pathogens. This was

supported by analysis of mRNA expression levels to find genes that are both essential and highly expressed at the mRNA level, suggesting that blocking translation will have significant effect on protein levels. A presentation to program leadership at DARPA's new headquarters resulted in an invitation to apply for the Phase II funding.

Poplar Xylome (http://xylome.vbi.vt.edu)

The DOE-funded project led by Eric Beers and Amy Brunner has resulted in growing the protein-protein interaction network surrounding wood development. While the project initially focused on a pre-defined binary screen using yeast-two-hybrid (Y2H) methods, supplemented with some library-screening, in this year, the research team has expanded that to pursue library screening with all interesting candidates, followed by binary validations. Library screening has proven to generate more hits than the pain-staking binary screen.

HortMutants

A new website (http://hortmutants.vbi.vt.edu) was developed to present ongoing insertion mutagenesis results in various plant species. First the strawberry mutants developed and published by Oosumi (2010) and Ruiz-Rojas (2010). Tomato mutants from former VBI faculty Andy Pereira, now at University of Arkansas. Also with Dr. Richard Veilleux of Horticulture and Barry Flinn of Forest Resources and Conservation, Dr. Dickerman supported student Sukhwinder Aulakh on the analysis of high-throughput Illumina mRNA-seq data on two very interesting mutants of potato. This analysis formed part of Aulakh's PhD thesis "Characterization of Activation Tagged Potato *(Solanum tuberosum L.)* Mutants" defended in September.

Data Analysis Core (DAC)

In November, Dr. Dickerman shifted part-time into the Data Analysis Core (DAC) of the Genomics Research Laboratory. Here, he provides analysis for large data in support of GRL customers. He works with Dr. Wyatt McMahon and Dr. Robert Settlage of the DAC and the new director of the Genomics Research Laboratory, Dr Saikumar Karyala. Dr. Dickerman's special focus in the DAC will be analysis of bacterial genome and metagenome data sets.

Publications

Shallom SJ, Tae H, Sarmento L, Preston D, McIver L, Franck C, Dickerman A, Adams LG, Garner HR. Comparison of genome diversity of Brucella spp. field isolates using Universal Bio-signature Detection Array and whole genome sequencing reveals limitations of current diagnostic methods. *Gene*. 2012 Nov1; 509(1): 142-8.

Veilleux RE, Mills KP, Baxter AJ, Upham KT, Ferguson TJ, Holt SH, Lu N, Ruiz-Rojas JJ, Pantazis CJ, Davis CM, Lindsay RC, Powell FL, Dan Y, Dickerman AW, Oosumi T, Shulaev V. Transposon tagging in diploid strawberry. *Plant Biotechnol J.* 2012 Oct; 10(8): 985-94.

Wattam AR, Inzana TJ, Williams KP, Mane SP, Shukla M, Almeida NF, Dickerman AW, Mason S, Moriyón I, O'Callaghan D, Whatmore AM, Sobral BW, Tiller RV, Hoffmaster AR, Frace MA, De Castro C, Molinaro A, Boyle SM, De BK, Setubal JC. Comparative genomics of early-diverging *Brucella* strains reveals a novel lipopolysaccharide biosynthesis pathway. *MBio*. 2012 Aug 28; 3(5): e00246-12.

HOESCHLE LABORATORY

The research in Dr. Ina Hoeschele's lab has two major components: collaborations with medical and biological scientists which require statistical design and analysis of genetical systems biology and genomics experiments, and methodology research which entails developing and comparing statistical methods for specific problems in genetics/genomics. Current genome-wide association studies (GWAS), including standard GWAS associating DNA variation with a small number of phenotypes, and human genomics studies such as our epigenomics project still rely on standard regression techniques to associate pairs of variables (e.g. single SNPs or single methylation sites with single transcripts or

disease phenotypes). This approach is conceptually flawed and underpowered. Dr. Hoeschle is therefore developing and testing alternative methods based on penalized regression and dimension reduction. Penalized regression methods have been developed for prediction and fail to perform well on large-scale variable selection problems such as these genomics datasets. Initial results on combining penalized regression with variable selection by false discovery rate control are encouraging. These methods are compared to Bayesian approaches.

Our most outstanding collaborative project in 2012 has been to provide statistical leadership and support to a large epigenomics project directed by Dr. Yongmei Liu at Wake Forest University Medical School. In this project, genome-wide profiling of DNA methylation (>485,000 CpG sites) and mRNA expression (>48K transcripts) was performed on monocyte samples from 1264 participants in the Multi-Ethnic Study of Atherosclerosis (MESA), a large cohort of subjects between the ages of 55 and 94 from four communities in the US. The team implemented a pre-processing pipeline for these data and contributed to a first study of the methylomics of gene expression in human monocytes which identified and characterized 11,203 potential cis-regulatory methylation sites. The current focus and the focus for the next year is to investigate the genetics of genome-wide gene expression and methylation in this population by using the available genome-wide SNP data (>3.8 million SNPs) in MESA. This information will help researchers understand the effects of genomic variation on DNA methylation and its influence on gene expression, and to quantify the mediating role of DNA methylation in the effects of genetic and non-genetic risk factors on disease.

Funded projects:

Haplotyping and QTL mapping in pedigrees with missing data. (Co-PI)
National Institutes of Health
1R01GM073766-01A2

Epigenome-wide association study of DNA methylation and artherosclerosis. (Sub-contract PI)
National Institutes of Health
1R01HL101250-01

Highly multivariate quantitative trait loci mapping in systems genetics. (PI)
National Institutes of Health
1R01HG005254-01

Genetic determinants of visceral adiposity: A genome-wide association study. (Sub-contract PI)
National Institutes of Health
5R01AG032098

Longitudinal study investigating the progression and pathogenesis of atypical hyperadrenocorticism in Scottish Terriers. (Co-PI)
AKC Canine Health Foundation

KALE LABORATORY

Research in the Kale lab is focused on infectious fungi of the human respiratory system. Research in the lab incorporates the use of comparative and functional genomics, bioinformatics, biochemistry, proteomics, and microbiology. *Aspergillus fumigatus,* an opportunistic fungus, causes a spectrum of disease phenotypes in humans, known broadly as Aspergillosis. In healthy individuals spores are inhaled and processed (killed and or cleared) through what are believed to be intrinsic cellular responses. However, in immunocompromised individuals spores are thought to germinate and then grow invasively throughout lung tissue. This phenomena, known as invasive aspergillosis, has high mortality if left untreated. Utilizing *Aspergillus fumigatus* as a model fungus that causes a breadth of disease lung phenotypes, the Kale lab hopes to abstract a systems level view of human invasive eukaryotic pathogens. There are three thematic areas regarding the study of invasive aspergillosis by the Kale lab.

Effectors

Microbe host interactions are mediated in part by small-secreted proteins, commonly known as effectors, that modulate host functions. A subset of these proteins act extracellularly while others

are able to translocate into cells. A number of commensal, pathogenic, and symbiotic bacteria deliver these effectors through needle-like apparatuses that directly pierce the host cells' plasma membranes. In 2010, Kale et al. showcased that a variety of eukaryotic plant pathogens utilize a completely different mechanism to deliver effector proteins into host cells by what is believed to be a functionally conserved N-terminal motif. These effectors trigger receptor-mediated endocytosis by binding cell surface phospholipids, specifically phosphatidylinositol-3-phosphate (PtdIns-3-P). Quite surprisingly, these effectors were shown to enter a number of mammalian cell types via PtdIns-3-P receptor- mediated endocytosis. Based on these novel findings, the Kale lab is determining if such effector proteins exist in the genomes of several medically relevant human fungal pathogens via comparative genomics and bioinformatics. A candidate list of putative effectors is currently being tested to determine cell entry and phospholipid binding activity. The discovery of effectors in mediating invasive fungal infection could potentially provide novel and new therapeutics, and challenge the paradigms for viewing disease development due to medically important fungi.

Endocytosis

The Kale lab is focused on understanding the mechanism of internalizing and processing spores associated with two prevalent cell types, airway epithelial cells and macrophages. Utilizing an array of techniques, the Kale lab has identified several general and specific markers associated with spore internalization. By understanding the mechanism of spore processing, researchers in the Kale lab are designing novel host-targeted therapeutics to bolster the innate immune response towards fungal spores.

Therapeutic Discovery and Development

The Kale lab has strong emphasis on translational sciences. The group has set up several i*n silico, in vivo*, and *in vitro* screens to identify and determine the efficacy of small molecules that block effector mediated entry. Blocking the delivery of effectors may potentially prevent or decrease the progression of disease and infection. Identification of such molecules could lead to the development of novel broad spectrum therapeutics against a growing list of eukaryotic pathogens that utilize such a mechanism to deliver effector proteins.

Publications

Kale SD. Oomycete and fungal effector entry, a microbial Trojan horse. *New Phytol.* 2012; 193: 874881. http://www.ncbi.nlm.nih.gov/pubmed/22403824

Kale SD, Tyler BM. Identification of lipid-binding effectors. *Methods Mol Biol.* 2012; 835: 393–414. http://www.ncbi.nlm.nih.gov/pubmed/22183667

Sun F, Kale SD, Azurmendi HF, Li D, Tyler BM, Capelluto DG. Structural basis for interactions of the *Phytophthora sojae* RxLR effector Avh5 with Phosphatidylinositol 3-phosphate and for host cell entry. *Mol Plant Microbe Interact.* 2012. http://www.ncbi.nlm.nih.gov/pubmed/23075041

Clark HR, Drews KC, Kale SD. Characterizing and measuring endocytosed lipid-binding effectors. *Methods in Enzymology: Endosomal Signaling.* In Press.

Activities

2012 NIH-Director's Early Independence Award Finalist.

Mentor - MAOP Fellowship 2012 (Yasir Hussein).

Mentor - Fralin Summer Fellowship 2012 (Kelly Drews).

Jackson Laboratories: Workshop on Surgical Techniques in the Laboratory Mouse.

Invited Seminars:

Functional characterization of the conserved modular domains found in the oomycete RXLR superfamily. American Phytopathology Society National Meeting. Providence, RI. August 2012.

The identification and characterization of RXLR-like effectors in medically relevant fungi. Virginia Bioinformatics Institute Research Symposium. Blacksburg, VA. March 2012.

Grant Support:

Illuminating the pathogen-host interface during invasive Aspergillosis.
Virginia Bioinformatics Institute and Fralin Life Sciences Institute's Small Grants Program

LAUBENBACHER LABORATORY

The focus of the Laubenbacher laboratory is on molecular systems biology, integrating the development and software implementation of theoretical tools and computational algorithms with applications to problems in biomedicine. The program is organized around three focus areas: the development of mathematical algorithms, their implementation into software tools, and the application of these tools to problems in systems biology.

The first focus area is currently funded by grants from the National Science Foundation and the US Department of Defense. The NSF-funded project is devoted to a computational study of the gene regulatory mechanism of canalization and the role of miRNAs. The DoD-funded project is devoted to the development of mathematical tools for optimal control of agent-based models, a popular model type for many complex systems.

The second focus area, software development, is funded by a grant from the National Science Foundation and includes a collaboration with three fungal immunology groups around the country. The goal of the project is to develop software for the construction, analysis, and use of computational models of molecular networks specifically tailored to use by experimentalists.

The third focus area, applications to biomedicine, is funded by two NIH grants, one from NCI, another from NIAID. Both projects have at their center mammalian intracellular iron metabolism. The NCI-funded project is in collaboration with researchers at the University of Connecticut Health Center and studies the role of iron metabolism in breast epithelial cells in the pathogenesis of breast cancer. The second project is in collaboration with the Lawrence laboratory at VBI and an immunology research group at the UVA School of Medicine. It studies iron metabolism as a component of the innate immune response to exposure to the common fungus *Aspergillus fumigatus*.

Publications

Hinkelmann F and Laubenbacher R. Finite fields in biology. In *Handbook of Finite Fields*. Mullen G and Panario D, eds. CRC Press, Boca Raton, FL, 2012. In press.

Murrugarra D, Veliz-Cuba A, Aguilar B, Arat S, and Laubenbacher R. Modeling stochasticity and variability in gene regulatory networks. *EURASIP J. Bioinf. and Sys. Biol*., 2012: 5.

Murrugarra D and Laubenbacher R. The number of multistate nested canalyzing functions. *Physica D,* 241, 929-938, 2012.

Chifman J, Kniss A, Neupane P, Williams I, Leung B, Mendes P, Hower V, Torti FM, Akman SA, Torti SV, and Laubenbacher R. The core control system of intracellular iron homeostasis: A mathematical model. *J. Theor. Biol*. 2012; 300, 91-99.

Veliz-Cuba A and Laubenbacher R, On the computation of fixed points in Boolean networks. *J. Appl. Math. Comp.* 2012; 39 (1-2),145-153.

Laubenbacher R. Mathematics in the public mind: The USA. In *Public Awareness of Mathematics*. Behrends E, Crato N, and Rodrigues F, eds. Springer Verlag, Heidelberg, 2012.

Presentations

Algebraic geometry in systems biology. Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH. January 20.

Patterns are everywhere: how and why? Kids' Tech University at Bowling Green State University, Bowling Green, OH. January 21.

Cancer Systems Biology. Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA. February 1.

Algebraic models in systems biology. Department of Mathematics, Mississippi State University, Mississippi State University, MS. February 16.

Algebraic geometry for large networks. DARPA Summit Meeting on Mathematics. Lake Tahoe, NV. February 21.

Algebraic models in systems biology. Workshop on Discrete and Topological Models in Molecular Biology. Tampa, FL. March 13.

Steady state analysis of Boolean network models via a universal class of models. Special Session on Dynamics of Complex Networks, AMS meeting. Washington, DC. March 18.

Wnt sigaling in melanoma cells. Special Session on Mathematical Methods in Disease Modeling, AMS meeting. Washington, DC. March 18.

Algebraic models in systems biology. Department. of Mathematics, University of Delaware, Newark, DE. April 2.

Algebraic models in systems biology. Department of Mathematics, University of Nebraska, Lincoln, NE. April 19.

Cancer systems biology. Department of Mathematics, University of Nebraska, Lincoln, NE. April 20.

Algebraic methods is systems and evolutionary biology. Mathematical Biosciences Institute. Columbus, OH. May 11.

Iron metabolism. Mathematical Biosciences Institute. Columbus, OH. May 18.

Science and the fight against cancer. Keynote address, American Cancer Society, Relay for Life. Tazewell, VA. May 31.

The role of SIAM as an advocate for the mathematical sciences community. SIAM Annual Meeting. Minneapolis, MN. July 12.

Downregulation of LRP6 inhibits growth of melanoma cells. SIAM Annual Meeting. Minneapolis, MN. July 13.

Algebraic models in systems biology. Department of Mathematics, Virginia Tech, Blacksburg, VA. September 21.

Algebraic models in systems biology. Oregon State University, Corvallis, OR. October 8.

Iron metabolism and the innate immune response to Aspergillus fumigatus. Oregon State University, Corvallis, OR. October 9.

Patterns are everywhere. Why and How? Kids' Tech University, Southwest Virginia Higher Education Center, Abingdon, VA.

Scientific posters: The good, the bad, and the ugly. Virginia Tech SIAM Student Chapter. November 13.

Funded projects

PlantSimLab: A simulation laboratory for plant biology.
National Science Foundation
DBI-1146819

The systems biology of iron homeostasis and the immune response to Aspergillus.
National Institute of Allergy and Infectious Disease-National Institutes of Health
1R21AI101619-01

A systems approach to iron metabolism in cancer cells.
National Cancer Institute-National Institutes of Health
1R21CA156133-01A1

Computational biomathematics: Toward optimal control of complex biological systems
US Army Research Office
W911NF0910538

Polynomial dynamical systems over finite fields: from structure to dynamics.
National Science Foundation
CMMI-0908201

REU Site: Modeling and Simulation in Systems Biology.
National Science Foundation
DMS-1062878

Integrated management of oomycete diseases of soybean and other crop plants.
USDA Agriculture and Food Research Initiative
Laubenbacher (co-PI) (PI: B. Tyler)

Service and Recognitions

Appointment to Fellow of the American Mathematical Society

Vice President for Science Policy, Society for Industrial and Applied Mathematics, 2009-2013

Member, Editorial Board, Journal of Algebra

Member, Editorial Board, Bulletin of Mathematical Biology

Member, Editorial Board, Journal of Symbolic Computation

Member, Editorial Board, Applied Mathematical Sciences book series, Springer Verlag

Member, Editorial Board, Mathematical Modelling: Theory and Applications book series, Springer Verlag

Member, Steering Committee, Visual Analytics in Biology Curriculum Network, Jackson State University, Jackson, MS, NSF funded, 2011-2013

Member, Scientific Advisory Committee, Mathematical Biosciences Institute, Ohio State University, 2011-2013

Member, Steering Committee for Activity Group in Algebraic Geometry, Society for Industrial and Applied Mathematics (SIAM)

CHRIS LAWRENCE LABORATORY

Research in the laboratory of Dr. Christopher Lawrence is focused primarily on pathogenic fungal-host interactions. In this regard two systems are being studied involving: 1) *Alternaria*, a fungus clinically associated with allergic airway disorders such a allergic rhinitis, asthma, chronic rhinosinusitis (CRS) and plant diseases, and 2) *Aspergillus fumigatus,* a fungus that causes deadly invasive aspergillosis and is particularly problematic in patients who are immunosuppressed or have other disorders such as cystic fibrosis.

This year as part of an NSF funded systematics project in collaboration with Barry Pryor (UAz) and Tobin Peever (Washington State), the Lawrence lab has completed the development of an Ensembl-based (http://www.ebi.ac.uk/), web accessible, database platform housing over ten fully annotated *Alternaria* genomes including both plant pathogenic and human allergy associated strains such as *A. alternata.* Along with an international consortium of researchers including scientists at the DOE-Joint Genome Institute, the Lawrence lab contributed to the comparative analysis of over 18 Dothideomycete fungal genomes. The *Alternaria* genomic sequences information and database platform at VBI has facilitated the identification and/or subsequent study of many genes in several species. In collaboration with Dr. Reinhard Fischer's group at the Karlsruhe Institute of Technology (KIT) in Germany, a polyketide synthase (PKS) gene responsible for the production of the mycotoxin Alternariol was discovered and characterized along with an entire suite of other PKS genes. Alternariol (AOH) has been shown to be an important deleterious contaminant of fruit and cereal products. Concern about exposure to low levels of AOH has been raised because contamination of food with the AOH-producing *Alternaria* species is associated with esophageal cancer. Moreover, AOH has been shown to have potent mutagenic properties in cultured cells. Efforts are now underway to deduce whether or not AOH is important in the context of inflammation in lung cells.

Genome sequencing coupled with functional genomics has also allowed the Lawrence lab in collaboration with Dr. Hirohito Kita, Director of Allergic Disease Laboratory at Mayo Clinic Rochester, Dr. Robert Cramer at Geisel School of Medicine at Dartmouth, Dr. Liwu Li at Virginia Tech, Dr. Brett Tyler at Oregon State University, and Dr. Shiv Kale at VBI to investigate the role of a secreted major allergen from *A. alternata* and study

its overall contribution to innate immunity and allergic inflammation. In this NIH-funded project it was discovered that an *Alternata* allergen has potent innate immunostimulatory activity in lung epithelial cells and drives Th2 inflammation in mouse models. In addition, this allergen has a RxLR sequence akin to a motif harbored by other secreted eukaryotic effector (virulence factor) proteins that facilitate entry into host cells. It was demonstrated this year that this motif in the allergen allows it to bind to phospatidylinositol 3-phospate (PI3P) and triggers endocytosis. Moreover, the lab has created allergen knockout mutants in the fungus allowing for studies at the whole organism level to ascertain the relative contribution of this allergen protein in driving immune responses *in vitro* and *in vivo*. The Lawrence lab is also collaborating with Mayo Clinic to ascertain the role of several other *Alternaria* proteins in airway inflammation. Lastly in regards to *Alternaria*, the Lawrence lab, in collaboration with Dr. Yangrae Cho at University of Hawaii, has functionally analyzed over one hundred transcription factors in the plant pathogenic species *A. brassicicola* and have identified several new plant virulence factors on Brassicas.

In regards to *A. fumigatus*, the Lawrence lab has established a NIH-funded collaboration with Dr. Laubenbacher at VBI to study the role of iron metabolism. In this project a systems biology approach is being taken to develop computational models of iron regulatory networks in both the fungus and in human cells such as macrophages and lung epithelial cells. The overall goal of this work is to identify new targets for therapeutic intervention in the context of deadly invasive fungal infections.

Finally, the Lawrence lab is studying redox signaling in plants and the role of the plant hormone jasmonic acid (JA) and its precursors. JA and its precursors/derivatives, collectively known as jasmonates, are a relatively new phytohormone class that regulate various aspects of plant developmental plasticity. In addition, their functions in stress acclimation signaling have been extensively investigated and have been established as key regulators of defense responses against pathogens and environmental stresses such as wounding, water deficit and ozone exposure. This year the Lawrence lab has discovered several "effector" proteins that physically bind to jasmonates in *Arabidopsis* plants using a biochemical approach. One of these effector proteins Cyp20-3, a cyclophilin, was shown this year to be critical for redox homeostasis via the formation of the cysteine synthase complex in chloroplasts and subsequent production of important sulfur containing secondary metabolites. Results of several recent studies have unveiled the broad-spectrum, selective anticancer activity of jasmonates. Now, the Lawrence lab members are utilizing their knowledge regarding how jasmonates signal in plants to understand the unknown mechanism underlying its unique anticancer activities in several cancers.

Manuscripts published in 2012:

Cho Y, Srivastava A, Ohm RA, et al. Transcription factor Amr1 induces melanin biosynthesis and suppresses virulence in *Alternaria brassicicola*. *PLOS Pathogens*. 2012; 8: e1002974.

Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T. Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. *BMC genomics*. 2012; 13: 171.

Ohm RA, Feau N, Henrissat B, et al. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLOS Pathogens*. 2012; 8: e1003037.

Saha D, Fetzner R, Burkhardt B, et al. Identification of a polyketide synthase required for alternariol (AOH) and alternariol-9-methyl ether (AME) formation in *Alternaria alternata*. *PLOS ONE*. 2012; 7: e40564.

Srivastava A, Ohm RA, Oxiles L, et al. A zinc-finger-family transcription factor, AbVf19, is required for the induction of a gene subset important for virulence in *Alternaria brassicicola*. *Molecular Plant-Microbe Interactions*. 2012; 25: 443–452.

Invited seminars 2012:

GBCB Graduate Program Seminar Series. Virginia Tech, Blacksburg, VA.

Geisel School of Medicine. Dartmouth College, Hannover, NH.

Research Live Series. Virginia Tech Carilion Research Institute. Roanoke, VA.

Funded Projects:

The role of the major *Alternaria* allergen Alt a 1 in airway inflammation.
National Institute of Health
NIAID 1R21AI094071-01

The systems biology of iron homeostasis and the immune response to *Aspergillus*.
National Institutes of Health
NIAID, 1R21AI101619-01

REU site: Modeling simulation in systems biology.
National Science Foundation
DMS-1062878

Alternaria and proteases in the development of Th2 immunity.
National Institutes of Health
NIAID R01AI71106-1A2

The pathogenesis of chronic rhinosinusitis.
National Institutes of Health
NIAID R01A1049235-09

Estimating speciation/reticulation boundaries in asexual *Alternaria* - a genomics approach.
National Science Foundation
DEB-0918298

Integrated management of oomycete diseases of soybean and other crop plants
NIFA-USDA
2011-68004-30104

MENDES LABORATORY

The Mendes laboratory creates computer models and simulations of biochemical networks, an activity that is at the core of systems biology. The group pursues two major areas: developing software for modeling and simulation and combining data analysis methods and modeling to infer networks from experimental data. The research is funded by the National Institutes of Health.

The software development activity is centered on the development and maintenance of the biochemical simulator COPASI (Complex Pathway Simulator) and related outreach activities. In the reporting period, one focus was on enhancing the usability of the software and adding the capability to specify validation data to parameter estimation tasks. The COPASI software continues being developed in collaboration with colleagues at the University of Heidelberg, Germany, and the University of Manchester, United Kingdom. The lab also developed a new application to speed up simulations using the distributed computing software Condor, a package known as Condor-COPASI.

Dr. Pedro Mendes collaborates with Dr. Reinhard Laubenbacher on reverse engineering methodology and on modeling iron metabolism. He also collaborates with several colleagues in the University of Manchester on various projects. Dr. Stefan Hoops collaborates with Dr. John Tyson's research lab in the Department of Biological Sciences at Virginia Tech in the development of COPASI-based software which provides user interfaces enabling COPASI to deal with rule-based models and hierarchical composed models.

Publications

Pahle J, Challenger JD, Mendes P and McKane AJ. Biochemical fluctuations, optimisation and the linear noise approximation. *BMC Systems Biology.* 2012; 6: 86.

Kent E, Hoops S, Mendes P. Condor-COPASI: High-throughput computing for biochemical networks. *BMC Systems Biology.* 2012; 6: 91.

Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, Mendes P, Swainston N. Improving metabolic flux predictions using absolute gene expression data *BMC Systems Biology.* 2012; 6: 73.

Heavner BD, Smallbone K, Barker B, Mendes P, Walker LP. Yeast 5 - an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Systems Biology.* 2012; 6:55.

Chifman J, Kniss A, Neupane P, Williams I, Leung, B, Deng Z, Mendes P, Hower V, Torti FM, Akman SA, Torti SV, Laubenbacher R. The core control system of intracellular iron homeostasis: A mathematical model. *J. Theoretical Biology.* 2012; 300, 91-99.

Davies O, Mendes P, Smallbone K, Malys N. Characterisation of multiple substrate-specific (d) ITP/(d)XTPase and modelling of deaminated purine nucleotide metabolism. *BMB Reports* 2012; 45, 259-264.

Funded projects

COPASI software for modeling and simulation of biochemical networks.
National Institutes of Health
R01-GM080219

Stochastic models of the cell cycle regulation in eukaryotes.
National Institutes of Health
R01-GM078989

REU site: Modeling and simulation in systems biology.
National Science Foundation
DMS-1062878

A simulation laboratory for plant biology.
National Science Foundation
DBI-1146819

BISWARUP MUKHOPADHYAY LAB

The Mukhopadhyay lab is investigating three redox systems in select archaea and bacteria. Two of these, heme-dependent sulfite reductases and thioredoxin based metabolic control, are being studied with the goal of understanding how these system evolved. The resulting knowledge will have broad applicability in the areas of biofuel production processes, mitigation of green house gas emission and human health. The third system concerns a deazaflavin coenzyme called coenzyme F_{420} and here the specific focus has been on how this coenzyme aids *Mycobacterium tuberculosis* (the bacterium that causes tuberculosis in humans) in evading as well as neutralizing bactericidal actions of host immune systems.

In the reporting period we have made many advances. A bioinformatic analysis has shown that the sulfite reductases of today's organisms likely originated from a family of heme-containing proteins of methane forming archaea or methanogens. These precursor heme proteins do not exhibit sulfite reductase activity and we call them DsrLP (dissimilatory sulfite reductase like protein). Methanogens utilize one of the most ancient respiratory metabolisms on Earth. Preliminary results suggest that methanogens of early Earth used DsrLPs for defense against oxidative damage and a study of these proteins could yield new information about heme biochemistry that will apply to organisms of all types. A combination of bioinformatic and experimental investigations showed that two well studied thioredoxin systems, NTR and FTR, were independently invented by the early archaea and bacteria, respectively, as tools for defense against oxidative damage and controlling carbon dioxide fixation. Work on coenzyme F_{420} metabolism in the mycobacteria have led to a new cell wall biosynthesis enzyme of *M. tuberculosis* that is the potential new target for the development of vaccines and therapeutics for TB **(Fig. 1).** A research collaboration with a vaccine manufacturer, PT Bio Farma in Indonesia, has been developed which will allow such discoveries with the mycobacteria to aid in production of a more effective TB vaccine. Collaborative work with DOE Joint Genome Institute and Russian Academy of Sciences, and many other laboratories in the US and abroad has led to the sequencing of the first cellulolytic archaeon, and the data show signs of novel metabolisms with respect to hydrogen production and cellulose biodegradation.

The VT-REU site, Microbiology in the post-genome era, has developed strong ties with several minority serving institutions for recruitment and student training. With renewed support from the NSF, the program continues to train undergraduate students from around the country in the summer,

and faculty laboratories from four colleges participate in this activity.

Evolution of a heme-based system – dissimilatory sulfite reductases

Hydrogen-dependent dissimilatory sulfate reduction ($4H_2 + SO_4^{2-} + H^+ \rightarrow HS^- + 4H_2O$) and hydrogenotrophic methanogenesis ($4H_2 + CO_2 \rightarrow CH_4 + 2H_2O$) are two of the most ancient energy-producing respiratory processes that developed at least 3.47 and 2.7-3.2 billion years ago, respectively **(Fig. 2)**. In general, methanogenesis and sulfate reduction are considered incompatible with each other because sulfite, an obligate intermediate of sulfate reduction ($SO_4^{2-} \rightarrow SO_3^{2-} \rightarrow HS^-$), inhibits methanogenesis. Dissimilatory sulfate reduction is performed by sulfate reducing archaea, and bacteria and biological methane formation or methanogenesis is a specialty of methane producing archaea called methanogens. Sulfite reductase, which removes sulfite in sulfate reducing organisms, has been thought be absent in methanogens. On the other hand, from an analysis of the geological data we have reasoned that methanogens of early Earth had to be sulfite tolerant and this ability continues to be important in the deep-sea volcanoes or hydrothermal vent environment that mimics some aspects of early Earth.

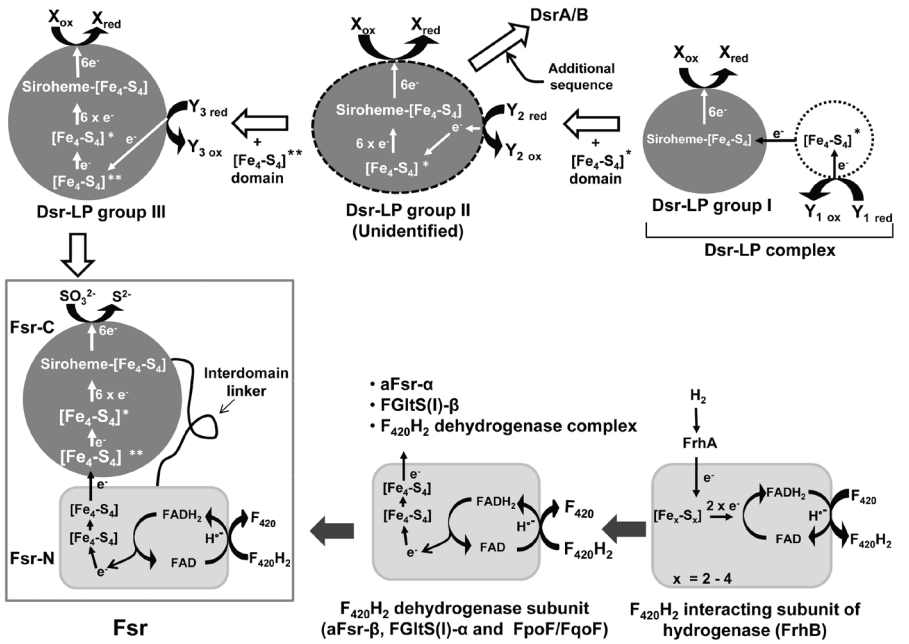


Figure 1. Development of dissimilatory sulfite reductase-like protein (Dsr-LP), $F_{420}H_2$-dependent sulfite reductase (Fsr) and dissimilatory sulfite reductase (Dsr). Fsr-N and Fsr-C: N-terminal and C-terminal halves of Fsr, respectively. FGltS(I)-α and FGltS(I)-β: $F_{420}H_2$ dehydrogenase and glutamate synthase subunit of a putative F420H2-dependent glutamate synthase of methanogens; FpoF/FqoF: $F_{420}H_2$ dehydrogenase subunit of a membrane-bound proton pumping $FF_{420}H_2$ dehydrogenase complexes of late evolving euryarchaea [36]; aFsr-α and aFsr-β: aSir and $FF_{420}H_2$ dehydrogenase subunits of a putative $FF_{420}H_2$-dependent assimilatory type siroheme sulfite reductase found in haloarchaea. * and ** are peripheral and additional iron sulfur clusters [Fe_4-S_4], respectively. Filled and unfilled boxed arrows show the path for the development of Fsr-N and Fsr-C, respectively. Dashed ovals or circles are unidentified proteins. X and Y_{1-3}, are unknown electron acceptors and donor, respectively.

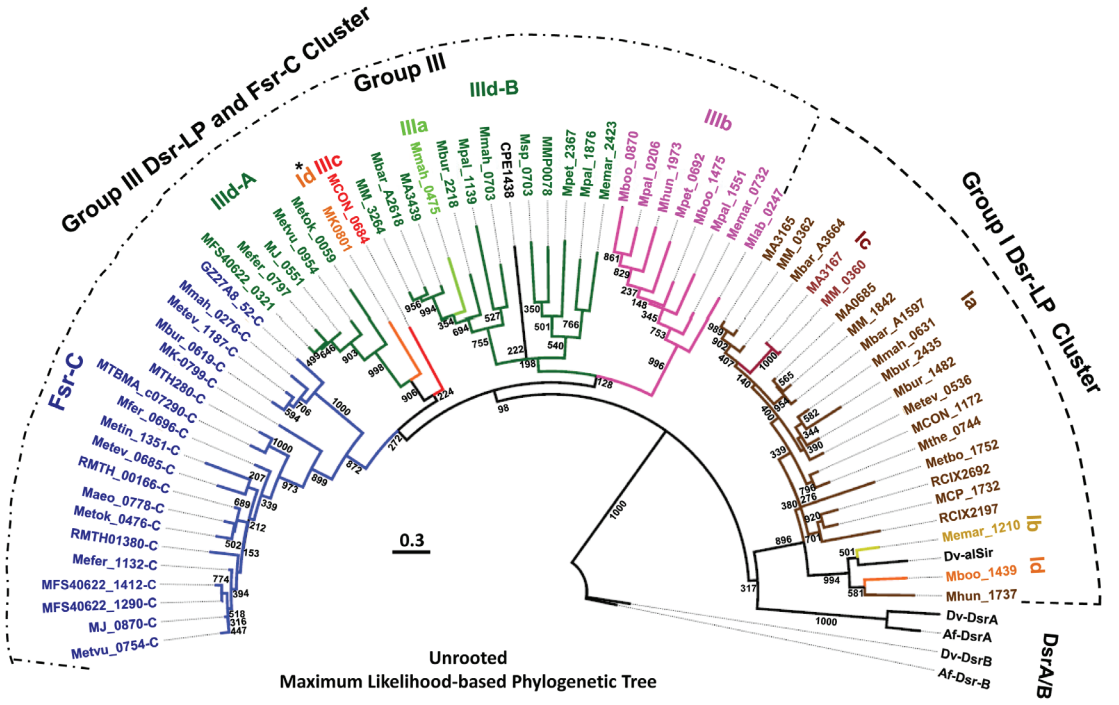


Figure 2. Phylogenetic tree of Dsr-LP and Dsr based on Maximum Likelihood method. Dsr-LP (Groups Ia-d and IIIa-d) and Dsr, Dsr and Fsr-C, defined in the legend of Fig. 1. Dv-DsrA/B and Af-DsrA/B, Dsr subunits A and B of *Desulfovibrio vulgaris* strain Hildenborough (ORFs DVU0402 and DVU0403) and *Archaeoglobus fulgidus* DSM 4304 (ORFs AF0423 and AF0424), respectively; Dv-alSir and CPE1438, anaerobic small sulfite reductase of *Desulfovibrio vulgaris* strain Hildenborough (ORF DVU_1597) and *Clostridium perfringens* strain 13, respectively; The ORF numbers followed by "-C", Fsr-C homologs. Abbreviation for organism names preceding the listed ORF numbers: MTBMA, *Methanothermobacter marburgensis* strain Marburg; MTH, *Methanothermobacter thermautotrophicus* ΔH; Metbo, *Mehanobacterium* sp. AL-21; RMTH, *Methanothermococcus thermolithotrophicus* (sequence obtained from Dr. William B. Whitman, University of Georgia); Mfer, *Methanothermus fervidus* DSM 2088; Maeo, *Methanococcus aeolicus* Nankai-3; Msp, *Methanosphaera stadtmanae* DSM 3091; MMP, *Methanococcus maripaludis* S2; Mpal, *Methanosphaerula palustris* E1-9c; Mpet, *Methanoplanus petrolearius* DSM 11571; Memar, *Methanoculleus marisnigri* JR1; MM, *Methanosarcina mazei* Gö1; MA, *Methanosarcina acetivorans* C2A; Mbar, *Methanosarcina barkeri* strain Fusaro; Mbur, *Methanococcoides burtonii* DSM 6242; Mmah, *Methanohalophilus mahii* DSM 5219; Metev, *Methanohalobium evestigatum* Z-7303; Mthe, *Methanosaeta thermophila* PT; MCON, *Methanosaeta concilii* GP-6; MCP, *Methanocella paludicola* SANAE; Mboo, *Candidiatus Methanoregula boonei* 6A8; Mlab, *Methanocorpusculum labreanum* Z; Mhun, *Methanospirillum hungatei* JF-1; Metbo, *Methanobacterium* sp. Al-21; Metok, *Methanothermococcus okinawensis* IH1; Metin, *Methanocaldococcus infernus* ME; MFS40622, *Methanocaldococcus* sp. FS406-22; MJ, *Methanocaldococcus jannaschii* DSM 2661; Mefer, *Methanocaldococcus fervens* AG86; Metvu, *Methanocaldococcus vulcanius* M7; MK, *Methanopyrus kandleri* AV19; GZ27A8_52, uncultured archaeon related to *Methanosarcina* and a member of an anaerobic methane oxidizing consortium; RCIX2692 and RCIX2197, uncultured methanogenic archaeon RC-1 and primary methane producer in rice rhizosphere. The bootstrap value shown at each branch is from 1000 replicates. Scale bar, number of amino acid substitutions per site. *, shows outliers. A and B, sub-clades of group IIId.

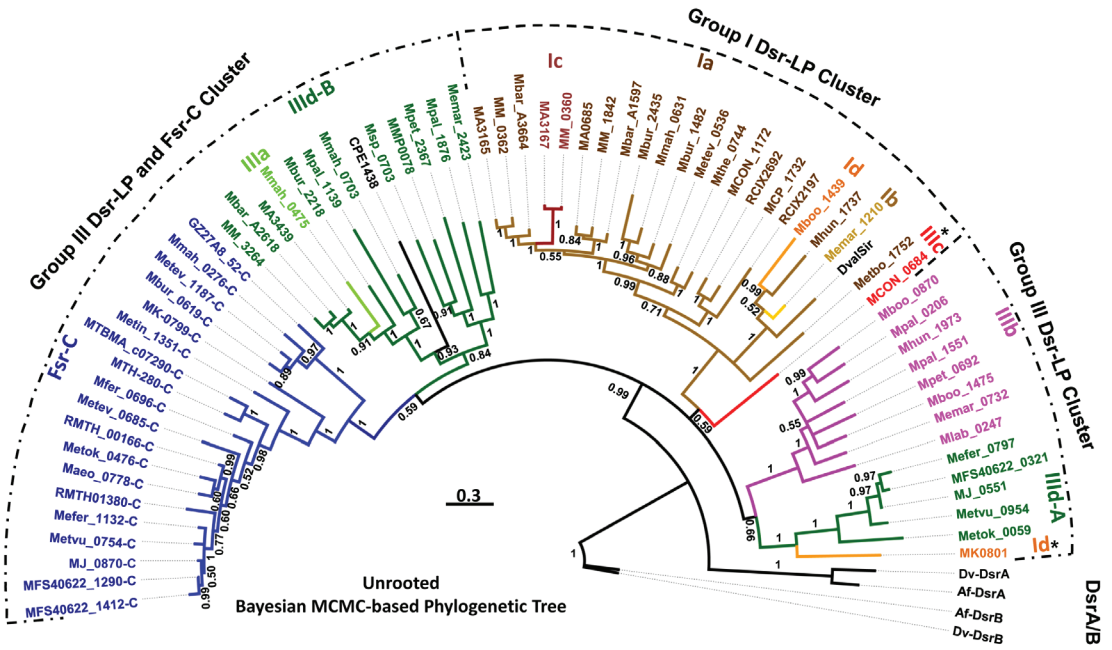


Figure 3. Phylogenetic tree of Dsr-LP homologs based on Bayesian Markov chain Monte Carlo (MCMC) analysis. ORF numbers and all abbreviations used are described in the legend of Fig. 2. For each branch a posterior probability value (0–1) is shown. Scale bar, number of amino acid substitutions per site.

Indeed, in the 2005-2012 period we showed that the methanogens that live in deep-sea volcanoes or hydrothermal vents metabolize sulfite and for this purpose they employ a special enzyme called F_{420}-dependent sulfite reductase (Fsr). Our data indicated that Fsr was probably built in methanogens from certain ancient parts. Starting from these clues, we have carried a bioinformatics based search and found that small heme-containing proteins of a variety of structures representing the essential core of sulfite reductases, is almost universally present in methanogens, although these proteins do not provide sulfite reduction activity to their host. We named these proteins dissimilatory sulfite reductase type protein or Dsr-LP. It is likely that Dsr-LP carries out a function that was essential to living cells on early Earth. The structural characteristics of the DsrLPs taken together provide a scheme for the evolution of heme-containing sulfite reductases in bacteria, plant and animal **(Fig. 3)**. Current research in the laboratory hints at the possibility of a new heme-based metabolism that could be widespread in living systems, including that of humans. This exploration has been funded by the NASA and a graduate fellowship from the GBCB PhD program.

Evolution of thioredoxin based redox control systems: Thioredoxin or Trx has been known as a master controller of critical metabolisms in a wide variety of living cells and the associated systems have been studied in detail in plant and animal. Some of the notable examples of Trx-controlled systems are photosynthesis in plants, and cell death and aging in humans. Trx systems have not been studied much in the anaerobic microorganisms some of which are the most ancient inhabitants of Earth. Our recent bioinformatics and experimental studies indicate that two branches of the Trx system, one (NTR) of which is dependent on a nicotinamide coenzyme and the other (FTR) on iron-containing proteins called ferredoxins, arose independently. The Mukhopadhyay lab found that FTR is likely an invention of ancient bacteria and NTR seemed to have been developed by the early archaea including methanogens (unpublished data). In both cases, the likely early utilities of the Trx systems have been in the control of carbon dioxide fixation and defense against oxidative damage. The latter aspect became an important metabolic function as oxygen appeared on Earth and the original inhabitants, which were accustomed to living without oxygen,

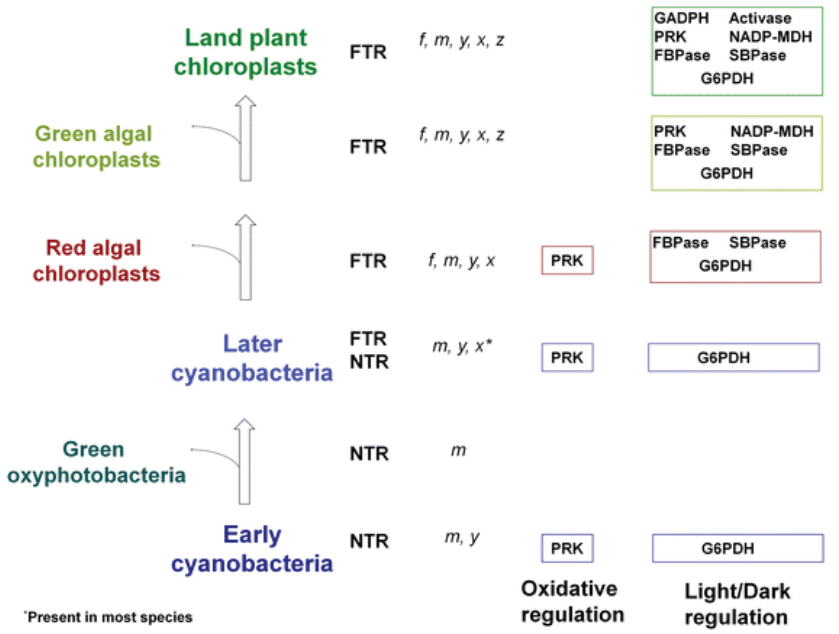


Figure 4. Scheme showing the evolutionary development of redox regulation in oxygenic photosynthesis. In plants, several enzymes of the Calvin–Benson cycle are subjected to classical light/dark regulation by the FTS (Schürmann and Buchanan 2008). The enzymes phosphoribulokinase (PRK), fructose 1,6-bisphosphatase (FBPase), sedoheptulose 1,7-bisphosphatase (SBPase), and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) are active upon reduction during light. In a coordinated manner, glucose 6-phosphate dehydrogenase (G6PDH), a key enzyme of the oxidative pentose-phosphate pathway, is light-deactivated. Other chloroplast enzymes, such as NADP-dependent malate dehydrogenase (NADP-MDH) and Rubisco activase, are also light-activated. During evolution, plants have developed an efficient network of metabolic control, and a clear transition from oxidative modulation-acquired from cyanobacteria-to a sophisticated light/dark-regulation in land plants. In addition, plants possess a large variety of Trxs in chloroplasts, namely, Trx-m, Trx-f, Trx-x, Trx-y, and Trx-z, that assure tight control and coordination of chloroplast metabolism with photosynthesis. In chloroplasts, FTR coexists with a special type of NTR (NTRC), of cyanobacterial origin, that exerts an important redox function in plants. While undergoing these changes, chloroplasts retained the oxidative type of regulation that was acquired by the parent cyanobacteria and that came to include a role for glutathione in the glutathionylation of certain enzymes.

had to deal with oxidative stress **(Fig. 3)**. Since the Trx system is an integral part of cellular redox-control, the new information resulting from the lab's research would have applicability in the production of biofuel and mitigation of greenhouse gas (methane) emissions and provide new leads for research on plant and animal metabolism. This research has been supported by a grant from the NSF and involves the following collaborators: Dr. Bob B. Buchanan, Department of Plant and Microbial Biology, University of California, Berkeley, CA; Dr. Mónica Balsera, Instituto de Recursos Naturales y Agrobiología de Salamanca (IRNASA-CSIC), Spain; Dr. Peter Schürmann, Laboratoire de Biologie Moléculaire et Cellulaire, Neuchâtel, Switzerland; and Dr. Ruth Schmitz-Streit, Institut für Allgemeine Mikrobiologie, Christian-Albrechts-Universität Kiel, Germany. The work on the evolution of FTR has been highlighted on the cover of an issue of Planta [Volume 237, Issue 2, February 2013, http://link.springer.com/journal/425/237/2/page/1].

Development of therapeutics and vaccine for tuberculosis

Coenzyme F_{420} is an essential component of the methanogenic archaea which is rarely found in bacteria and is absent in humans. Mycobacteria, which include *Mycobacterium tuberculosis*, the causative agent of tuberculosis or TB, are a rare group of bacteria that carries F_{420}. Ongoing research indicates that F_{420} could play a critical role in TB pathogenesis. More recently, the Mukhopadhyay lab has found that some of the coenzyme F_{420}-dependent proteins help to build the resilient cell wall of *M. tuberculosis*, and they may occur on the cell surface. Consequently, these proteins are attractive targets for the development of TB vaccines and drugs. To capitalize on this opportunity, Dr. Mukhopadhyay has developed a collaboration with PT Bio Farma, an Indonesian company producing international quality vaccines that are used worldwide, including the USA. The goal is to engineer a strain of *Mycobacterium bovis* BCG that is currently used as a tuberculosis vaccine (live and attenuated) at Bio Farma to carry new protein antigens that would be suitable as a TB vaccine. To facilitate this collaboration, Virginia Tech and PT Bio Farma have recently signed an MOU. The MOU provides opportunities of broader collaboration with other Virginia Tech research groups for research and education and outreach activities.

Genomics of the first known cellulolytic archaea, a hyperthermophile from Kamchatka peninsula of Russian Federation

A high temperature process for the conversion of cellulose to hydrogen is expected to provide higher conversion rates, reduced chances of contamination, energy savings (as heating a bioreactor is more efficient than cooling), and better access to bound cellulose within natural substrates. Thus far, most research on such studies has focused on bacterial and eukaryotic systems, as archaea were not known to degrade cellulose. In 2005, Dr. Elizaveta Bonch-Osmolovskaya and one of her graduate students, Dr. Anna A. Perevalova, at the Winogradsky Institute of Microbiology of the Russian Academy of Sciences, reported the isolation of the first cellulolytic archaeon, *Desulfurococcus fermentans*, from hot springs at the Uzon Caldera in the Kamchatka peninsula of Russia. *D. fermentans* is a hyperthermophilic archaeon, belonging to the phylum of Crenarchaeota. Interestingly, whereas other hydrogen producing anaerobes, including other Desulfurococcus species, are inhibited by accumulating hydrogen, *D. fermentans* continues to produce this product without inhibition. Thus, *D. fermentans* could harbor new cellular machineries and employ new mechanisms for cellulose degradation and hydrogen production, and these can be leveraged for industrial production of hydrogen from renewable resources.

To capitalize on this opportunity, the Mukhopadhyay laboratory has initiated a collaborative effort towards deciphering the molecular mechanism for uninhibited hydrogen production by *Desulfurococcus fermentans* from cellulosic material. In collaboration with Dr. Bonch-Osmolovskaya and many other researchers from the US and Japan, a comparative genome analysis project was initiated in which the features of *D. fermentans* are being compared with that of other *Desulfurococcus* species. This will provide the genomic basis for all molecular studies that follow. This collaborative effort received competitive support from the DOE Joint Genome Institute (DOE-JGI) for the sequencing of *D. fermentans* and two other *Desulfurococcus* species. The first paper from this study, led by the Mukhopadhyay lab has now been published. To help move this project from genome data to detailed phenotypic analysis, Dr. Perevalova spent seven months in the Mukhopadhyay laboratory in 2012 (January-August) conducting studies on the potentially new cellulolytic systems of *Desulfurococcus fermentans* in collaboration with two of our PhD students, Dwi Susanti and Jason Rodriguez. The results were presented by Dr. Perevalova in a poster at the 9th International Congress on Extremophiles, Sevilla, Spain, September 10-13, 2012.

NSF funded REU site: Microbiology in the post-genome era

This site was established in 2009 for training undergraduate students and high school teachers in microbiology research with approaches of the post-genome era. In 2012 NSF renewed the funding for its continuation for another three years. The site has

about 30 mentors who represent almost all colleges at Virginia Tech.

The site has several special features. The students manage enrichment activities by: serving on committees with guidance from the director whenever needed, visiting relevant industries, attending weekly seminars by VT faculty members, attending seminars by external speakers, visiting various laboratories on campus, organizing midsummer and end of the program symposia, joining regular lunch meetings with faculty mentors and graduate students, and sharing in a weekly informal discussion session. Students prepare a paper about their research targeted to a non-science audience and share it with family and friends. They also interact with younger students (junior high or high school students) through organized activities. In the summer of 2012, the lab trained a total of 15 students (one Hispanic/Latino, four African Americans, 11 females). Recruitment of talented and deserving students was facilitated through collaboration with the previous participants of the site and the following units: the Department of Biology and the Howard Hughes Medical Institute via the Comprehensive Integrated Biomedical Science Initiative Program at Oakwood University (Huntsville, AL); the Department of Chemistry and Native American Center at Fort Lewis College (Durango, CO); the Native American Research Laboratory at University of Montana, and the Department of Biology, University of Minnesota-Morris.

Short courses

NextGen genomics for microbial pathogenesis research. Institut Teknologi Bandung, Bandung, Indonesia. September 2012.

Workshops

Mukhopadhyay B, Purwantini E, Torto-Alalibo G, Kyrpides N. 2012. 5th MENGO workshop: Gene ontology terms for energy-related microbial genomes. Walnut Creek Marriott, Walnut Creek, CA. March 23, 2012. [Followed the 2012 DOE Joint Genome Institute (JGI) User Meeting and facilitated by the JGI].

Publications

Susanti D, Johnson EF, Rodriguez JR, Anderson I, Perevalova AA, Kyrpides N, Lucas S, Han J, Lapidus A, Cheng J-F, Goodwin L, Pitluck S, Mavromatis K, Peters L, Land ML, Hauser L, Gopalan G, Chan PP, Lowe TM, Atomi H, Bonch-Osmolovskaya EA, Woyke T, and Mukhopadhyay B. Complete genome sequence of *Desulfurococcus fermentans*, a hyperthermophilic cellulolytic crenarchaeon isolated from a freshwater hot spring in Kamchatka, Russia. *Journal of Bacteriology* 2012; 94: 5703–5704.

Susanti D and Mukhopadhyay B. An intertwined evolutionary history of methanogenic archaea and sulfate reduction. *PLOS ONE*. 2012; 7: e45313.

Invited Seminars

Hydrothermal vent biology of *Methanocaldococcus jannaschi* and development of sulfate reduction pathway. SRI International, Menlo Park, CA. June 2012.

Development of biological redox control and sulfate reduction systems. Indonesian Society for Microbiology, Institut Teknologi Bandung, Bandung, Indonesia. September 2012.

Development of biological sulfate reduction and redox control systems - from early Earth to today's Earth. School of Life Sciences, University of Nevada. Las Vegas, NV. October 2012.

Development of biological sulfate reduction and redox control systems - from early Earth to today's Earth. DuPont Central Research Development. Wilmington, DE. November 2012.

Poster presentations

Susanti D, Wong JH, DeSantis R, Vensel WH, Tanaka CK, Hurkman WJ, Loganathan U, Schmitz-Streit R, Balsera M, Buchanan BB, and Mukhopadhyay B. Thioredoxin-linked redox control of metabolism in an evolutionarily deeply-rooted hyperthermophilic methanogen. General Meeting of the American Society of Microbiology. San Francisco, CA. June 16-19, 2012.

Perevalova A, Steen I, Stokke R, Susanti D, Bonch-Osmolovskaya E, Mukhopadhyay B. Molecular basis for polysaccharides degradation capabilities of *Desulfurococcus fermentans*, a hyperthermophilic crenarchaeon. 9th International Congress on Extremophiles. Sevilla, Spain. September 10-13, 2012.

Purwantini E, Torto-Alalibo G, Setubal J, Tyler B, Mukhopadhyay B. Microbial energy processes gene ontology (MENGO): Development of terms and annotation for energy-related microbial genomes. 2012.

Purwantini E, Torto-Alalibo G, Setubal J, Tyler B, Mukhopadhyay B. Gene ontology terms describe biological production of methane. US Department of Energy Genomic Science Awardee Meeting X. Bethesda, MD. February 26-29, 2012.

Purwantini E, Torto-Alalibo G, Setubal J, Tyler B, Mukhopadhyay B. Gene ontology terms describe biological production of methane. Seventh Annual US Department of Energy Joint Genome Institute, User Meeting, Walnut Creek, CA. March 20-22, 2012.

Purwantini E and Mukhopadhyay B. A coenzyme F420-dependent enzyme generates ketomycolic acids in *Mycobacterium tuberculosis* and it is a new target for a tuberculosis drug PA-824. General Meeting of the American Society of Microbiology, San Francisco, CA. June 16-19, 2012.

Torto-Alalibo G, Purwantini E, Setubal J, Tyler B, Mukhopadhyay B. Microbial energy processes gene ontology (MENGO): New gene ontology terms describing microbial processes relevant for bioenergy. US Department of Energy Genomic Science Awardee Meeting X. Bethesda, MD. February 26-29, 2012.

Torto-Alalibo G, Purwantini E, Setubal J, Tyler B, Mukhopadhyay B. Microbial energy processes gene ontology (MENGO): New gene ontology terms describing microbial processes relevant for bioenergy. Seventh Annual US Department of Energy Joint Genome Institute User Meeting. Walnut Creek, CA. March 20-22, 2012.

SETUBAL LABORATORY

Research in the laboratory of Dr. João Setubal is focused on the analysis of microbial genome, transcriptome, and metagenome sequence data and on the development of computational tools for such analysis. Current work on bacterial genomes includes *Liberibacter americanus*, *Xylella fastidiosa*, and *Xanthomonas citri*, all of which are citrus pathogens; *Mycobacterium abcessus,* an opportunistic human pathogen; strains in the genus *Gluconacetobacter,* which are capable of synthesizing cellulose and hence have promising biotechnological potential; and *Azotobacter vinelandii*, a model organism for the study of nitrogen fixation and iron-sulfur clusters. The lab is now devoting considerable efforts to the São Paulo Zoo metagenomics project, which aims to study a variety of microbiota, in particular those that play a role in composting processes. The lab also has a project which studies retinopathy of prematurity (a condition in premature babies that can cause blindness) using mouse models and RNA-seq transcriptome data (collaboration with R. Giordano).

All of these projects deal with vast amounts of data created by next generation sequencing instruments. Dr. Setubal's lab applies a combination of computational tools and develops its own processing pipelines. One current theme of interest is phylogenomic networks, which allow the study of gene flow in the biosphere and therefore have the potential of providing insights into the emergence of new pathogens (in plant, animals, and humans). Given the gigantic amount of sequence data currently available and expected to become available in the next few years, building and analysis of these networks requires special techniques and algorithms that the Setubal group is currently developing.

Publications

Wilson ML, Cai Y, Hanlon R, Taylor S, Chevreux B, Setubal JC, Tyler BM, Peccoud J. Sequence verification of synthetic DNA by assembly of sequencing reads. *Nucleic Acids Research*. 2012; doi: 10.1093/nar/gks908

Wattam A, Inzana T, Williams K, Mane S, Shukla M, Almeida N, Dickerman A, Mason S, Moriyon I, O'Callaghan D, Whatmore A, Sobral B, Tiller R, Hoffmaster A, Frace M, De Castro C, Molinaro A, Boyle S, De B, and Setubal JC. Comparative genomics of early-diverging *Brucella* strains reveals a novel lipopolysaccharide biosynthesis pathway. *MBio*. 2012; 3(5): e00246-12.

Saldanha H, Ribeiro E, Borges C, Araujo A, Gallon R, Holanda M, Walter ME, Togawa R, and Setubal JC. Towards a hybrid federated cloud platform to efficiently execute bioinformatics workflows. In *Bioinformatics*. InTech Publishers, 2012.

Yang K, Heath L, and Setubal JC. REGEN: Ancestral genome reconstruction for bacteria. *Genes*. 2012; 3(3), 423-443.

Dias Z, Dias U, Setubal JC. SIS: a program to generate draft genome sequence scaffolds for prokaryotes. *BMC Bioinformatics*. 2012; 13(1):96.

Dos Santos PC, Fang Z, Mason SW, Setubal JC and Dixon R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics*. 2012; 13(1):162.

Paschoal AR, Maracaja-Coutinho V, Setubal JC, Paulino Simões ZL, Verjovski-Almeida S and Durham AM. Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases. *RNA Biology*. 2012; 9(3).

Collen PN, Collen J, Reis MS, Pedersen M, Setubal JC, Varani AM, Colepicolo P and Oliveira MC. Analysis of expressed sequence tags from the agarophyte *Gracilaria tenuistipitata* (*Rhodophyta*). *Journal of Applied Phycology*. 2012; 24(4):641-647.

Warren A and Setubal JC. Using entropy estimates for DAG-based ontologies. In *Proceedings of the 15th Bio-Ontologies Special Interest Group Meeting of ISMB*. 2012.

Dias Z, Dias U, Heath L, and Setubal JC. Sorting genomes using almost-symmetric inversions. In *Proceedings of the 27th Symposium on Applied Computing*. March 26-30, 2012, Riva del Garda, Italy.

Invited seminars:

Center of Biotechnology, Federal University of Rio Grande do Sul, Brazil.

Department of Biochemistry, Federal University of Minas Gerais, Brazil.

Department of Biophysics, Federal University of São Paulo, Brazil.

Conference lectures:

First Brazilian Genome Conference. Foz do Iguaçu, Brazil.

X-Meeting 2012. Campinas, Brazil (Keynote lecturer).

Fourth Research Integration Forum. Federal University of São Paulo, Brazil.

Workshop on Bioinformatics and Algorithms, University of São Paulo, Brazil.

Conference technical program committee memberships:

2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICAABS). Las Vegas, NV.

International Symposium on Bioinformatics Research and Applications (ISBRA). Dallas, TX.

Brazilian Symposium on Bioinformatics (BSB). Campo Grande, Brazil.

23rd International Conference on Genome Informatics. Taiwan.

Bioinformatics. Portugal.

Professional recognition

Elected Association of Computing Machinery (USA) Distinguished Speaker

Funded projects:

Studies of microbial diversity in the São Paulo Zoo.
FAPESP (Brazil)

Development of protein computational models for bioenergy relevant microbes.
FAPESP (Brazil)

Director, Genome Sciences Research Center. University of São Paulo, Brazil.


CID
CYBERINFRASTRUCTURE

SUMMARY

The CyberInfrastructure Division (CID) has been pursuing integrated approaches and bioinformatics infrastructure development for the analysis and determination of countermeasures for infectious diseases. The award of the bacterial Bioinformatics Resource Center, PATRIC, and the Pathogen Portal to CID as a competitive renewal of the original PATRIC project provides the foundation for the evolution of CID's methodologies and distributed computational infrastructure. These projects drove our change to a JBoss portal website development platform, inclusion of the RAST genome annotation system (through a subcontract to the University of Chicago and Argonne National Laboratory), and a production-level website and database with failover capability. The Pathogen Portal serves as a gateway into all the Bioinformatics Resource Centers or BRCs (viral, influenza-specific, bacterial, eukaryotic pathogens, and vectors) and provides programmatic access to proteomics, gene expression, and other data types for the BRCs, which is coordinated through an Interoperability Working Group led by CID. CID's approach to developing bioinformatics resources is user-centered, engaging users early to understand their research questions (use cases) and later to evaluate emerging prototypes. This approach helps ensure that the resources we develop have high utility and usability in the research communities they are intended to serve. In addition, CID continues to use its model of extensive collaborations with diverse scientific communities across the globe to jointly discover new biological knowledge, as well as focusing internally on biological research that CID's infrastructure, data and analysis are uniquely positioned to develop. Some select examples are shown for key host-associated bacteria and progress in bioinformatics infrastructure development and deployment.

BIOINFORMATICS INFRASTRUCTURE DEVELOPMENT FOR INFECTIOUS DISEASE

PATRIC Bacterial Bioinformatics Center

The Pathosystems Resource Integration Center (PATRIC, patricbrc.vbi.vt.edu) is a web-based information system designed to support basic and applied biomedical research on bacterial infectious diseases. Funded by the National Institute of Allergy and Infectious Diseases (NIAID) as a Bioinformatics Resource Center (BRC) for bacteria, PATRIC integrates genome-scale data, metadata, and analysis tools for bacterial pathogens and other bacteria. PATRIC focuses on NIAID Category A-C bacterial pathogens, but provides data, tools, and analysis services for all publicly available bacterial genomes. As of November 2012, PATRIC has released the genomic data for a total of 7622 bacterial genomes. To enable comprehensive comparative analyses, genome annotation in PATRIC is performed in a standardized manner using the RAST (Rapid Annotation using Subsystem Technology) system. RAST predicts genes, assigns gene functions, and reconstructs metabolic pathways. It is powered by a robust assembly of subsystems that have been curated based on evaluation of hundreds of prokaryotic genomes and the clustering of common protein families encoded within these genomes (FIGfams). PATRIC also provides a free end-user genome annotation service through RAST to allow users to annotate their own genomes. Genome metadata parsed from genome project data and curated from other sources supports searching for and locating genomes of interest based on various combinations of 61 different metadata fields (e.g., host, phylogentic relationship, lifestyle, isolation location, etc.) PATRIC updates its genomic data, including addition of newly available genomes, on a monthly basis.

In addition to the RAST annotations, PATRIC includes other reference annotations and an Identifier (ID) mapping tool, which allows users to quickly map between PATRIC annotation identifiers and identifiers used by various resources, such as PDB, RefSeq, etc. PATRIC's Protein Family Sorter **(Fig. 1)** compares protein families across closely related or diverse groups of genomes, visualizes them using interactive heatmaps and generates multiple sequence alignments and phylogenetic

trees for individual families. The heatmap view is an interactive visualization tool that provides an overview of the distribution of proteins across a selected set of genomes.

The Comparative Pathway Tool comparison of consistently annotated metabolic pathways across closely related or diverse groups of genomes and visualizes them using interactive KEGG maps and heatmaps (providing an overview of the distribution of genomes across the set of EC numbers within a selected pathway). The Disease View integrates infectious disease, host, pathogen and disease outbreak data, which enables infectious disease-centric access and analysis of host-pathogen interactions. The Phylogeny Viewer allows exploration of phylogenetic relationships using species- and genus-level coloring schemes and supports custom creation of genome groups for use in other PATRIC tools.

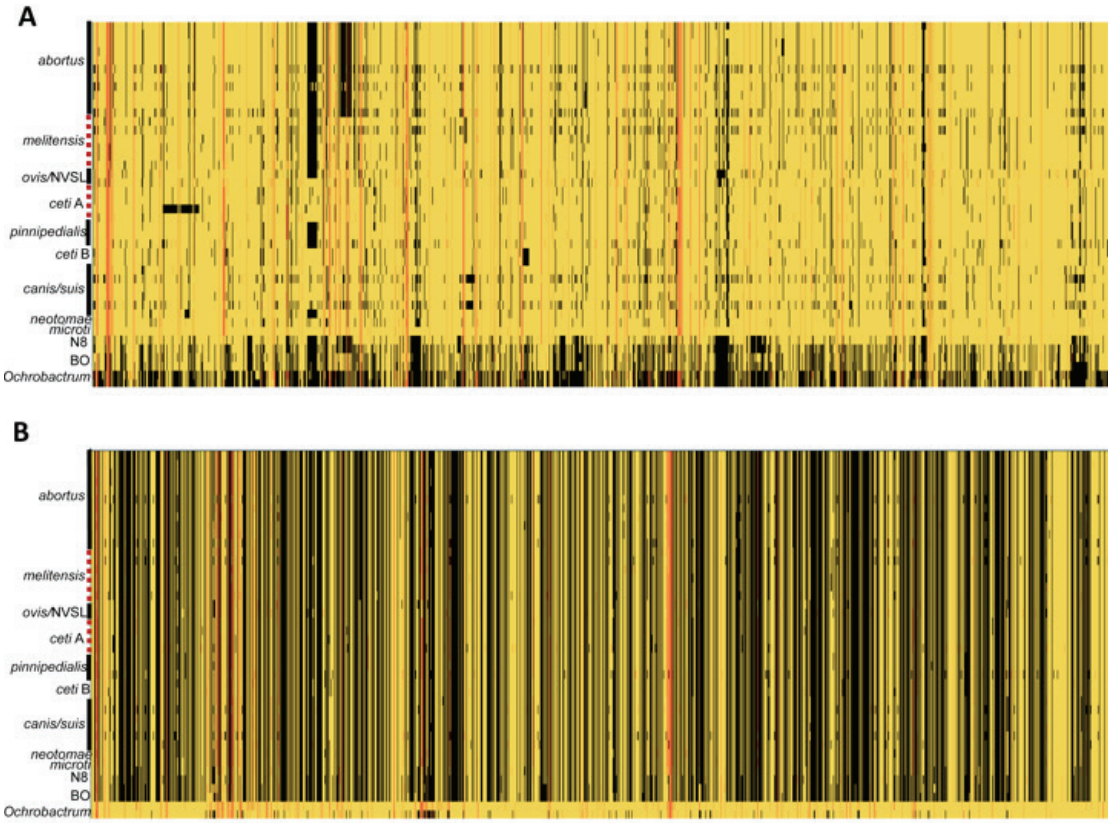


Figure 1. The first half of the brucellaceae pan proteome with protein families oriented by either *B. microti* (A) or *O. anthropi* (B).

In response to user feedback and requests, we have deployed the PATRIC Workspace wherein researchers can register and create their own private account in PATRIC to save, manage and analyze the data groups gathered from the PATRIC site. To further aid researchers in accessing and using the PATRIC data and tools, we undertook a major website redesign **(Fig. 2)**. New features include: a completely redesigned homepage showcasing biologically-relevant workflows that illustrate how users can accomplish important tasks using PATRIC; direct access to PATRIC's BLAST interface allowing users to copy-and-paste a nucleotide or amino acid sequence and BLAST against all PATRIC data or against specialty subsets such as plasmids and transcriptomics data; a visual mosaic of a

subset of PATRIC's most powerful tools with brief descriptions; and a revamped tag cloud depicting the most recently-accessed bacterial genomes at PATRIC as measured using live Google Analytics website traffic statistics.

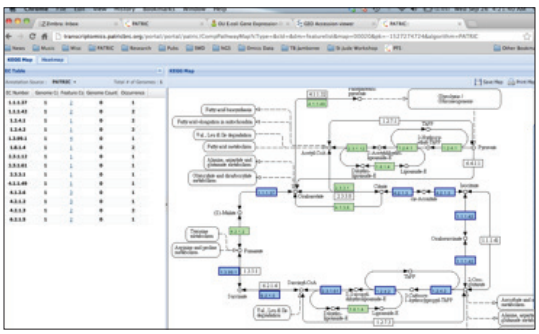
The workflows in particular allow users to get a quick understanding of how to use specific PATRIC pages and tools within the context of a biologically-relevant question. Each workflow includes detailed PATRIC screen images and instructive text so that users can conduct similar workflows with data relevant to their datasets and/or interests. Lastly, each workflow contains a number of links to other areas within PATRIC related to either tools and/or data pages illustrated in the workflow. Current workflows include finding genomic islands, identifying proteins from outbreak strains, and comparing diverse biochemical pathways. We will be adding new workflow pages on a regular basis as we grow PATRIC's suite of data and capabilities.

Figure 2. PATRIC's updated website design. The new website design includes workflow pages, a consolidated tool landing page, and improved site-wide look and feel to enhance PATRIC's overall usability.

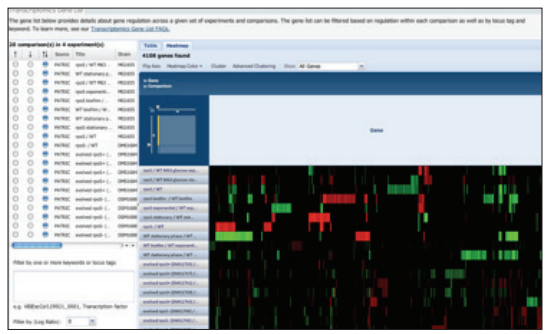
Most recently, PATRIC has released a suite of integrated methods and tools to explore, visualize, analyze and compare a large number of published gene expression data sets available at PATRIC, as well as upload and analyze unpublished gene expression data sets, including RNA-Seq data. PATRIC has curated and incorporated 465 gene expression datasets with 3249 comparisons for bacterial pathogens from NCBI's GEO database. Organism and gene identifiers described in the expression array platform are mapped to corresponding genomes and genes in PATRIC. Data from replicates are merged, normalized and log-transformed using a manual curation process for quality control. Experimental procedures and sample metadata are also curated to accurately and consistently capture information such as sample strain, genetic modification, experimental condition, treatment, and time-point. Then, the expression data are combined with the other genomic data in PATRIC to provide integrated data analysis capabilities. For any taxonomy level or genome, the available data sets are displayed as experiment and sample lists with metadata-based searching and progressive filtering capability. For each of the samples, numbers of differentially expressed genes identified to be significant ($|\log \text{ratio}| \geq 1$ or $|Z\text{-score}| \geq 2$) are summarized and linked to the corresponding gene list. Multiple data sets can be selected for further analysis across multiple

experiments and/or can be saved as a group for repeated or future use. Researchers can create a gene list based on selected datasets, displaying all the genes reported in the selected data sets, their functions and expression level summaries.

Gene lists can be dynamically filtered to find the most differentially expressed genes by applying different log ratio and/or Z-score thresholds. Lists can also be filtered by gene names or functions to analyze expression of any genes of interest. For a gene of interest, highly positively and negatively correlated gene lists are displayed along with their function. This allows researchers to identify potentially co-regulated genes or genes that perform similar functions and, often, generate hypothesis about potential function of a hypothetical gene. Some of the expression data-related tools and functionality available on the PATRIC website are shown in **Fig. 3A-D** below.



A) Pathway enrichment



B) Heatmap with clustering

C) Comparisons and all conditions

D) Comparisons and top conditions

Figure 3. Examples of PATRIC gene expression data analysis tools and visualizations.
A) Pathway enrichment - allows researchers to quickly find top metabolic pathways corresponding to their genes of interest and visualize them in KEGG (Kyoto Encyclopedia of Genes and Genomes) maps along with all other genes annotated in a pathway. **B)** Heatmap and clustering - shows expression levels of all the genes in a gene list across all selected samples. Genes can be filtered and sorted based on their genomic locations and using hierarchical clustering. Any area of the heatmap can be selected to download corresponding expression data or to save the genes as a group. **C & D)** For a given gene, top-ranking samples in which the gene expression passes the specified threshold are listed along with their metadata, such as sample strain, genetic modification, and experimental conditions along with a visual summary of the metadata.

PATRIC also allows researchers to upload unpublished gene expression data into their private workspace to explore them with various analysis and visualization tools and to compare them with the public datasets at PATRIC. Pre-processed gene expression data generated by using either microarray or high-throughput sequencing technologies can be uploaded to PATRIC as Excel or tab-delimited files in the form of a gene list or a gene matrix. An additional file containing sample metadata is also provided to aid in the data analysis. Transcriptomics data generated using high-throughput sequencing technologies can be first processed using the RNA-Seq Analysis Pipeline available at Pathogen Portal (see section below) and then be imported to the PATRIC workspace for further analysis **(Fig. 4)**.

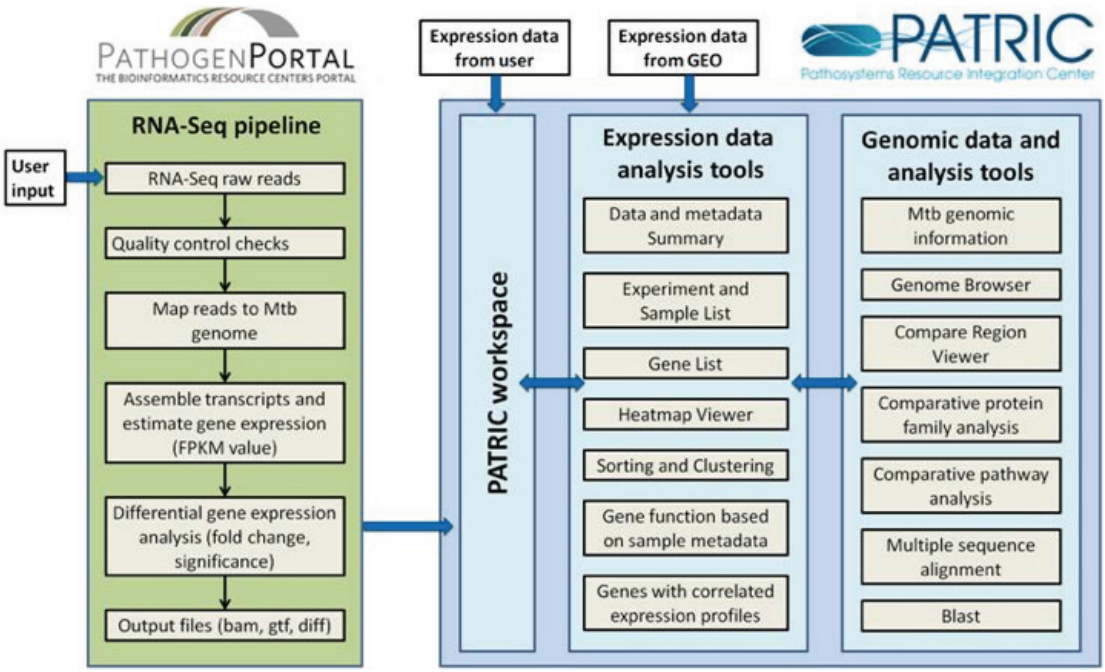


Figure 4. RNA-Seq analysis workflow in Pathogen Portal and PATRIC.

RNA-SEQ ANALYSIS PIPELINE AT PATHOGEN PORTAL

Pathogen Portal (www.pathogenportal.org) is focused scientifically on enabling comparative analysis of host response to pathogens, while providing common infrastructure services across BRCs when appropriate. Pathogen Portal has recently developed an RNA-Seq pipeline with the goal of providing a guided experience for processing and analyzing high-throughput sequencing data to characterize the transcriptome of pathogens, their hosts and other related organisms. This resource provides analysis capability for human, mouse, rat and zebrafish hosts in addition to 3729 different genomes from the PATRIC, VectorBase and EuPathDB BRCs. Genomes for use as references by the RNA-Seq pipeline are frequently updated as new sequences are published at the BRCs. The pipeline runs on Amazon's EC2, and is free to use. Built on the Galaxy open source bioinformatics workflow system infrastructure, Pathogen Portal's RNA-Seq pipeline provides alignment and quantitative expression data; records provenance information, including the tools and parameters used to process data; supports batch analysis for multiple samples; and provides results for sharing and publishing.

The Galaxy system has been modified by the Pathogen Portal team to help simplify the process of RNA-Seq analysis for routine use by informatics-

naïve, biologically-focused users and provide a guided experience to quality control of read data, read mapping, assembling transcripts, estimating gene expression values, and doing differential expression analysis **(Fig. 4)**. Researchers can upload read data into their own private project space at the Pathogen Portal, analyze and trim their reads using base-call quality, and use the system to contrast expression profiles for various conditions. The resulting data can also be used to discover new genes and alternative transcription start sites. The RNA-Seq Pipeline is available at http://rnaseq.pathogenportal.org.

SCIENTIFIC PROGRESS

COMPARATIVE STUDIES AND COLLABORATIONS INVOLVING PATHOGENIC BACTERIAL GENERA

Advances in Rickettsial biology and a search for bacterial sequences in eukaryotic genomes

Several members of PATRIC (Gillespie, Driscoll, Nordberg, Sobral) have participated in the community project for the genome sequencing of the deer tick (*Ixodes scapularis*), an important arthropod vector of many pathogenic microbial species. Our collaboration with the J. Craig Venter Institute and other members of the community involved an in-depth analysis of a rickettsial species (Alphaproteobacteria: Rickettsiales: *Rickettsiaceae*) that was serendipitously sequenced along with the tick genome. Importantly, this genome from the *Rickettsia* endosymbiont of *Ixodes scapularis* (REIS) is the first genome from *Rickettsia* spp. to be assembled entirely from a eukaryotic sequencing project, as the bacterial reads were separated from the *I. scapularis* whole genome shotgun sequencing data. Aside from this novelty, the characteristics of this rickettsial genome were extraordinary in comparison to other sequenced *Rickettsia* genomes, and most importantly, highlighted a hereto undescribed look into a dynamic rickettsial mobilome that carries many important genes for obligate intracellular survival, and mechanisms of pathogenicity by comparing this symbiont genome to genomes of closely related human pathogens.

Genome-based phylogeny estimation suggests that REIS is the most basal member of the spotted fever group (SFG), sharing attributes with both pathogenic and non-pathogenic rickettsiae **(Fig. 5A)**. REIS is the largest *Rickettsia* genome sequenced to date (>2Mb), with a chromosome size >1.8Mb that contains 2059 predicted open reading frames (ORFs) **(Fig. 5B).** The accessory genome alone encodes almost twice as many genes as the entire *R. prowazekii* genome. Plasmids pREIS1-4 are novel to *Rickettsia* species, and contain 59, 83, 43, and 65 predicted ORFs, respectively. Many of these ORFs have homologs in other species of diverse obligate intracellular bacteria. Perhaps the most outstanding finding within the REIS genome is the extraordinary proliferation of mobile genetic elements (MGEs), which contributes to a limited synteny with other *Rickettsia* genomes and places REIS among the most repetitive bacterial genomes. In particular, a Rickettsiales amplified genetic element (RAGE), previously identified as a "proliferated-yet-decayed" element in scrub typhus rickettsiae (*Orientia tsutsugamushi*) genomes, is present on both the REIS chromosome and plasmids (pREIS1 and pREIS3). However, unlike *O. tsutsugamushi*, nine complete (or nearly complete) RAGEs exist within the REIS genome and encode entire F-like type IV secretion systems similar to the single copy tra cluster present in the genomes of *R. bellii* and *R. massiliae*. An unparalleled abundance of encoded transposases (>650) relative to genome size, together with the RAGEs and other MGEs, comprise >35% of the total genome, making REIS one of the most dynamic bacterial genomes sequenced to date.

Despite an extraordinarily plastic accessory genome, including several intriguing lateral gene transfers (e.g., a complete biotin biosynthesis operon, remnants of the WO-B prophage, a Gram-positive aminoglycoside antibiotic synthesis and

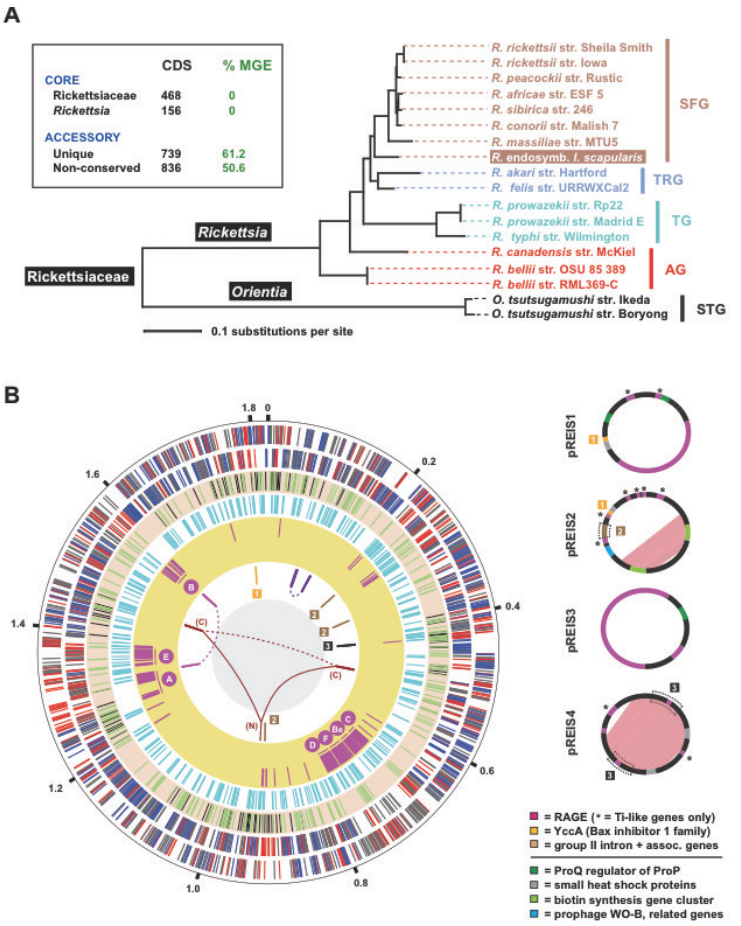


Figure 5. Phylogenomic analysis of the genome sequence of the Rickettsiales endosymbiont of *Ixodes scapularis* (REIS). A) Phylogeny estimated from 191 concatenated protein alignments. STG = scrub typhus group, AG = ancestral group, TG = typhus group, TRG = transitional group, SFG = spotted fever group. Inset illustrates brief results of orthologous group (OG) clustering of 2309 total predicted REIS ORFs across 16 Rickettsiaceae genomes. In green, the percentage of mobile genetic elements per OG category is provided. **B)** Characteristics of the REIS genome. (Left, chromosome) Putative origin of replication, positioned at 12 o'clock. Outer black circle is a scale with coordinates (in Mb) listed at 200-kb intervals. Six rings inside the scale are as follows: (1) plus strand and (2) minus strand genes with core *Rickettsia* genes colored blue, REIS singletons (not present in other *Rickettsiaceae* genomes) colored red, and all other genes colored gray; (3) 232 putative pseudogenes colored green and gaps (108 total) in the assembly colored black; (4) mobile genetic elements including transposases, integrases, phage related ORFs, and other genes typically encoded on plasmids colored turquoise; (5) Rickettsiales amplified genetic element (RAGE) genes colored pink, with the location of the seven complete or nearly complete RAGE clusters illustrated (C, Be, F, D, A, E, and B); (6) regions of large scale duplication and recombination (joined by lines), and regions resulting from plasmid integration (boxed numbers 1-3). Ring-six color scheme as follows (clockwise from origin): purple, duplication of five ORFs and the 6S RNA gene; brown, group II intron and associated genes; black, six gene insert from pREIS4; burgundy, N-terminal region of *rickA* (N) joined to the duplicated C-terminal regions (C); magenta, four gene duplication associated with RAGE-A and RAGE-B. (Right, plasmids) Main features of the four REIS plasmids. Duplicated regions within pREIS2 and pREIS4 are depicted with pink shading. Multi-gene regions (described above) that have been transferred to the chromosome are within dashed boxes. Color scheme (bottom right) depicts genes present on both the chromosome and plasmids (above line) and other noteworthy features of the plasmids.

resistance cluster) from other diverse intracellular bacteria, comparative analysis with 15 *Rickettsiaceae* genomes indicates that REIS does not differ in the mode and tempo of gene loss typical of rickettsial reductive genome evolution. As REIS is not known to invade vertebrate cells and has no known pathogenic effects on *I. scapularis*, its genome sequence provides an invaluable tool for deciphering the mechanisms of arthropod and vertebrate pathogenicity via comparison with the genomes of virulent and avirulent *Rickettsia* species. Of significance to *Rickettsiaceae* evolution, we present compelling evidence that many of the genes involved in obligate intracellular lifestyle were acquired via MGEs, especially the RAGE, through a continuum of genomic invasions by these integrative conjugative elements.

At the time of its publication, the REIS genome was the sixteenth *Rickettsia* genome sequenced to date, with its four distinct plasmids (pREIS1-4) bringing the total number of sequenced rickettsial plasmids to eleven. Remarkably, an explosion of sequenced rickettsial genomes has occurred over the last year, bringing the total number of *Rickettsia* genomes to almost fifty. Our research team has analyzed most of these genomes in a recent contribution to the American Society of Microbiology book titled *Intracellular Pathogens II: Rickettsiales*. Despite these additional genomes, as well as comparison across 80 Rickettsiales genomes, our results for REIS still stand and accentuate the outstanding nature of this genome.

This success of our project highlights the great opportunity for further studying a variety of bacterial genomes that have never been cultured or isolated. We are currently pursuing the extraction and assembly of other partial (or nearly complete) bacterial genomes from a variety of eukaryotic genomes. The goal is to incorporate this information into PATRIC, and to cross-reference (whenever possible) extracted data with their respective host genomes available at other BRCs. We have very recently completed work on a partial rickettsial genome extracted from an animal genome (the placozoan *Trichoplax adhaerens*), and submitted this paper in November 2012.

BRUCELLA PHYLOGENOMICS: NEW STRAINS, NEW PHYLOGENIES, AND NEW TREATMENT APPROACHES.

Since 2009, we have been in collaboration with three groups on the comparative analysis of *Brucella* genomes. One such collaboration has already resulted in a publication (Wattam, Williams et al. 2009) as well as an invitation to submit a book chapter (Sobral and Wattam, 2011).

Our collaboration with Barun De of the Centers for Disease Control (CDC) has concentrated on four new genomes, two from human infections, and two closely related genomes that were isolated from Australian rodents. The bioinformatics analysis led to the discovery that one of the new strains that had been isolated from a human patient was missing many of the genes important in forming the O-antigen of the lipopolysaccharide (LPS), a virulence factor for *Brucella* **(Fig. 6)**. In place of the usual genes, the BO2 strain had four different genes known to produce a different type of LPS in other bacteria. Verification of the alternative LPS structure was verified experimentally with two new collaborators, Tom Inzana from the Virginia-Maryland Regional College of Veterinary Medicine, and Antonio Molinaro from the Department of Organic Chemistry and Biochemistry, University of Naples Federico II, Naples, Italy. The manuscript, "Comparative genomics of early-diverging *Brucella* strains reveals a novel lipopolysaccharide biosynthesis pathway," has been published in MBio.
A separate collaboration with Stephen Boyle and Nammalwar Sriranganathan, both from the Virginia-Maryland Regional College of Veterinary Medicine, has identified a novel method to limit the growth of *Brucella* that have successfully invaded macrophages. Peptide nucleic acids (PNAs) linked to cell penetrating peptides (CPP) have been shown to penetrate the bacterial cell envelope, inhibit expression of targeted genes and cause inhibition

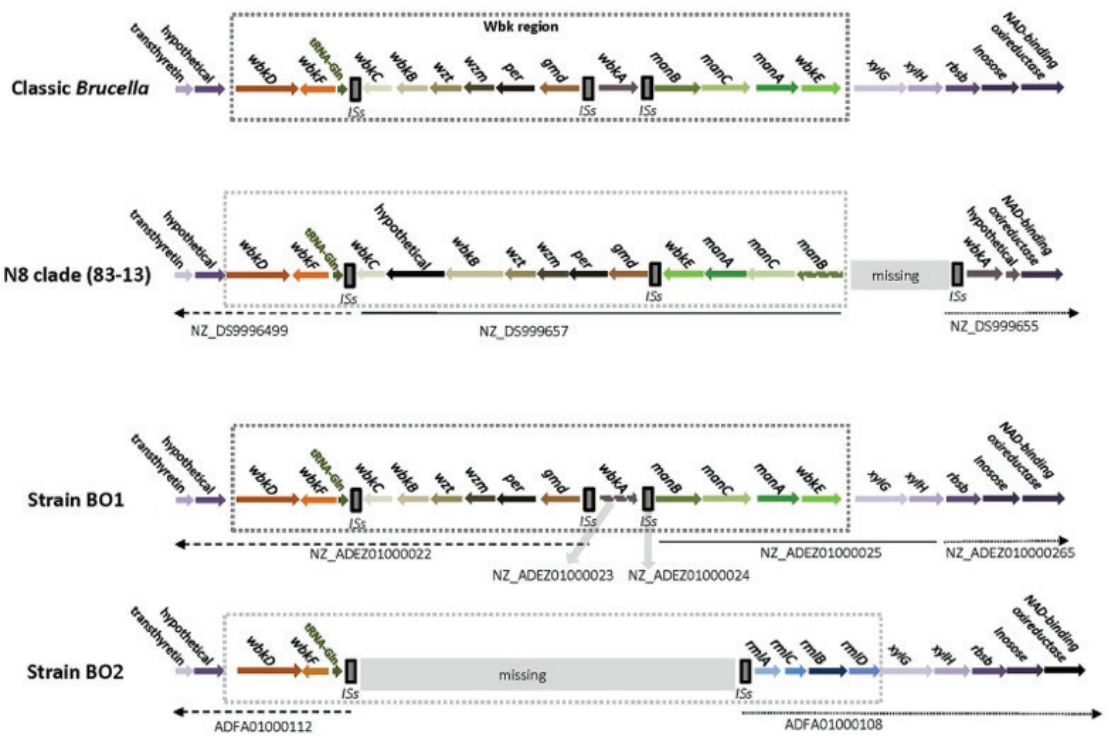


Figure 6. Comparison of the wbk regions across *Brucella* strains. The wbk region contains most of the genes important in O-antigen synthesis in *Brucella*, an important determinant in virulence.

of growth of both Gram-positive and -negative bacteria. This study showed that some of the PNAs targeted to *Brucella* genes involved in DNA (polA, dnaG, gyrA), RNA (rpoB), cell envelope (asd), fatty acid (kdtA, acpP) and protein synthesis (tsf) inhibited the growth of *B. suis* in culture and in macrophages after 24 hours of treatment. A manuscript, "Peptide nucleic acids inhibit *Brucella suis* in pure culture and infected macrophages," has been accepted at the International Journal of Antimicrobial Agents.

A long-standing collaboration between CID and scientists at Northern Arizona University, Universite Montpellier, University of California, the Broad Institute and the Animal Health and Veterinary Laboratories Agency in the UK examined 40 *Brucella* genomes. This analysis has produced a manuscript currently under review at MBio, "Out of the dirt and into vertebrates: Comparative phylogenomics and evolution of the *Brucellae*," that documents the evolution of the *Brucellae*, identifying the specific steps that were most vital in adapting from free-living soil bacterium into an intracellular niche **(Fig. 7)**. Extensive genome reduction accompanied by lateral transfer events allowed the stepwise acquisition of virulence factors, including a type IV secretion system, a perosamine-based O-antigen and systems for sequestering metal ions absent in progenitors. Subsequent explosive radiation within the core *Brucella* resulted in lineages that appear to have evolved within their preferred mammalian hosts, toning down their virulence to become stealth pathogens capable of causing long-term chronic infections.

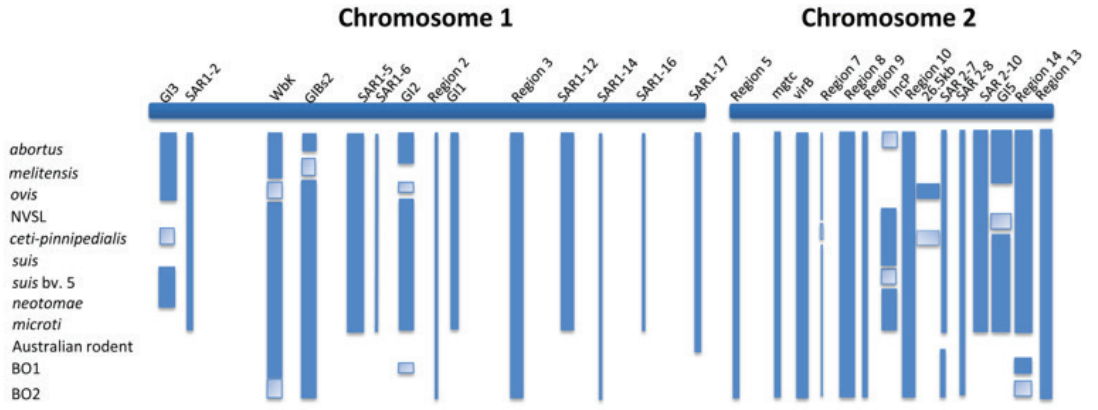


Figure 7. Location of genomic regions of interest across a "typical" *Brucella* of two chromosomes. Dark blue bands indicate that a genome or clade has all the genes in the region present. A lighter band indicates that a genome or clade is missing some or all of the genes in a region. Absence of a band shows that the genome or clade does not have this region.

INTERACTIONS WITH THE MYCOBACTERIUM RESEARCH COMMUNITY

Mycobacterium tuberculosis (Mtb) is a pathogenic bacterium and the causative agent of tuberculosis (TB). TB affects two billion people and causes nearly two million deaths yearly worldwide. To facilitate TB research, the genome of the *Mycobacterium tuberculosis* H37Rv strain was sequenced and annotated in 1998. Recently, the genome of Mtb H37Rv was re-sequenced by several sequencing centers to improve the genome sequence accuracy and genome annotation. Several annotation sources published the genome annotations of H37Rv: RefSeq (Pruitt and Maglott 2001), Broad Institute / Tuberculosis Database (TBDB) (Reddy, Riley et al. 2009), PATRIC, and Tuberculist (Lew, Kapopoulou et al. 2011). **Table 1** shows the CDS annotations from RefSeq, Broad, and PATRIC. It also compares the number of hypothetical proteins predicted from each of the annotation sources.

Combining all different annotations, there are still 508 proteins annotated as "hypothetical proteins" by all three sources. Hundreds of putative enzymes from *Mtb* remain categorized as "conserved hypothetical proteins" or "hypothetical proteins," offering little or no information on their functional role in pathogenic and non-pathogenic functions. To improve the protein annotation, CID members (Mao, Shukla, and Sobral) collaborated with Dr. Luiz Pedro de Carvalho's group at MRC National Institute for Medical Research and used fold recognition and three-dimensional structure prediction as an alternative approach for annotation of protein-coding genes. The advantage of this type of approach is the unbiased search of homologous proteins, which might have very low primary sequence identity, but fold in the same way.

Table 1. Comparison of CDS annotations and hypothetical proteins from RefSeq, Broad and PATRIC (March 2012).

Annotation	RefSeq (Original H37Rv genome)	Broad (H37Rv Broad genome)	PATRIC (H37Rv Broad genome)	Shared identical or equivalent CDSs*
Total CDSs	3989	4160	4301	3812
Hypothetical proteins	1481	2795	1034	508

*: The identical or equivalent CDSs shared among all three sources.

In this study, the fold and structure of nearly all proteins encoded in the genome of *Mtb* H37Rv using Phyre[2] (Protein Homology/analogY Recognition Engine V2.0) was predicted. A comparison of the Phyre2 predicted protein functions with the annotations from three different sources (RefSeq, Broad and PATRIC) was performed. Out of 3339 proteins with Phyre2 confidence scores of 75% and above, 2975 matched the template proteins with known functions. An examination of the corresponding RefSeq, Broad, and PATRIC annotations of these 2975 proteins found that 836 from RefSeq, 1765 from Broad, 325 from PATRIC, and 242 from all three sources were annotated as "hypothetical" **(Fig. 8)**. This result suggests that the structure prediction can facilitate annotation of hypothetical or unknown gene products. The structure prediction not only helps predict protein functions of those tagged as "hypothetical" from the genome annotations, but also verifies annotations with "non-hypothetical" or "known" proteins. Many examples where Phyre[2] predictions verified some of the PATRIC annotations when Broad and RefSeq annotated as "hypothetical" and verified some of the Broad annotations when PATRIC and RefSeq annotated as "hypothetical" were found. The results are summarized in a submitted manuscript ("Functional assignment of *Mycobacterium tuberculosis* proteome revealed by genome-scale fold-recognition" by Mao, Shukla, et al.). The analysis indicated that a fold-recognition and structural homology might be used as a complementary tool in genome annotation efforts; furthermore, it can deliver primary sequence-independent information regarding structure, ligands, and even substrate specificity for enzymes that display low primary sequence identity.

PATRIC co-hosted the Tuberculosis Community Annotation Project (TBCAP) Jamboree with the Broad Institute GSCID, Tuberculosis Database (TBDB), and NIAID in March 2012. The purpose of the jamboree was to engage members of the TB research community to improve the annotation of the *Mycobacterium tuberculosis* genome sequence and increase the value of this community resource. The jamboree served as a forum for community members to contribute data and knowledge that

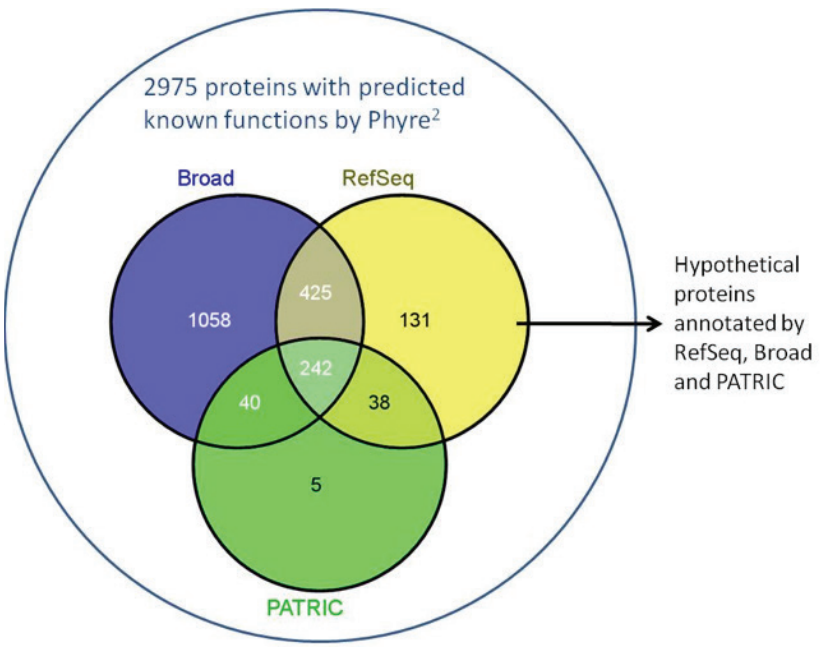


Figure 8. Hypothetical proteins from different genome annotation sources that overlap with proteins with predicted known functions by Phyre[2].

will be captured and displayed with respect to the genome sequence, and to provide the TBCAP community members an opportunity to review and revise annotations before they are publicly released. As an outcome of the jamboree, several collaborative manuscripts are in development that will be released in a special issue of the journal Tuberculosis.

Another follow-up to the TB Annotation Jamboree was a collaboration with Dr. William Jacobs, Jr., at the Albert Einstein College of Medicine and the Howard Hughes Medical Institute. This collaboration has already produced a publication submitted to Nucleic Acids Research about plasmid transformation in *Mycobacterium smegmatis*, "The Efficient Plasmid Transformation Phenotype of *Mycobacterium smegmatis* mc2155 is mediated by the loss of a chromosomal binding protein." We are also working with Drs. Jacobs and Joann Tufariello on the comparative genomics of a new genome, *M. haemophilum*, a close relative to *M. leprae*.

CID BIOINFORMATICS WORKSHOPS

Members of CID have been invited to give bioinformatics workshops that feature PATRIC. Individual workshops are listed below.

Texas A&M University, College Station, TX. Host: Jim Samuels. February 2012. 53 participants.

TB Annotation Jamboree. Virginia Tech, Arlington, VA. Hosts: PATRIC, Broad Institute, TBDB. March 2012. Over 20 participants.

Albert Einstein College of Medicine, Bronx, NY. Host: Bill Jacobs Jr. June 2012. 29 participants.

Duke University and University of North Carolina, Durham NC. Hosts: David Tobin and Miriam Baurnsten. October 2012. Over 50 participants.

Web-based workshop for University of California, San Diego and international collaborators. Host. Don Catanzaro. November 2012. 13 participants.

EXTERNAL PRESENTATIONS

Sobral, BW. "H37RV Genome Annotation Jamboree and PATRIC and Pathogen Portal BRCs." Many Hosts of Mycobacteria: Advancing Translational Science. NIAID, Bethesda, MD. July 2012.

Sobral, BW. "InformaticsEnabled MicrobeHost-Environment Interactions." 20th Annual International Conference on Intelligent Systems for Molecular Biology, ISMB. Long Beach, CA. July 2012.

Wattam R. "PATRIC: A Resource for infectious disease research." New York State Department of Health, Wadsworth Center. Albany, NY. September 2012.

Wattam, R. "PATRIC: A Resource for infectious disease research." St. Jude's Children's Research Hospital. Memphis, TN. September 2012.

Stevens R and Will R. "PATRIC BRC." 4th Annual Systems Biology Programmatic Meeting. Richland, WA. November 2012.

Wattam, R. "A practical overview of PATRIC for Brucella researchers." International Brucellosis Conference. Chicago, IL. Dec 2012.

PUBLICATIONS

Gillespie J, Joardar V, Williams KP, Driscoll T, Hostetler JB, Nordberg E, Shukla M, Walenz B, Hill CA, Nene VM, Azad AF, Sobral BW and Caler E. A *Rickettsia* genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle. *J Bacteriol* 2012. **194**(2): 376-394.

Gillespie, JJ, Nordberg E, Azad AF, Sobral BW. Phylogeny and comparative genomics: The shifting landscape in the genomics era. *Intracellular Pathogens II: Rickettsiales*. Azad AF, Palmer GH, eds. American Society of Microbiology. 2012.

Jirimutu Z, Wang G, Ding G, Chen Y, Sun Z, Sun H, Zhang L, Wang S, Hasi Y, Zhang J, Li Y, Shi Z, Xu C, He S, Yu S, Li W, Zhang M, Batmunkh B, Ts N, Unierhu S, Bat-Ireedui H, Gao B, Baysgalan Q, Li Z, Jia T, Subudenggerile N, Wang Z, Wang J, Pan L, Chen Y, Ganerdene Y, Dabxilt E, Altansha A, Liu T, Cao M, Aruuntsever B, Hosblig, He F, Zha-Ti A, Zheng G, Qiu F, Sun Z, Zhao L, Zhao W, Liu B, Li C, Chen Y, Tang X, Guo C, Liu W, Ming L, Temuulen A, Cui Y, Li J, Gao J, Li W, Niu S, Sun T, Zhai Z, Zhang M, Chen C, Baldan T, Bayaer T, Li Y and Meng H. Genome sequences of wild and domestic bactrian camels. *Nat Commun.* 2012; 3: 1202.

Kappmeyer LS, Thiagarajan M, Herndon DR, Ramsay JD, Caler E, Djikeng A, Gillespie JJ, Lau AO, Roalson EH, Silva JC, Silva MG, Suarez CE, Ueti MW, Nene VM, Mealey RH, Knowles DP and Brayton KA. Comparative genomic analysis and phylogenetic position of *Theileria equi*. *BMC Genomics.* 2012; **13**(1): 603.

Kaur SJ, Rahman MS, Ammerman NC, Beier-Sexton M, Ceraul SM, Gillespie JJ and Azad AF. TolC-dependent secretion of an ankyrin repeat-containing protein of *Rickettsia typhi*. *J Bacteriol.* 2012; **194**(18): 4920-4932.

Pyysalo S, Ohta T, Rak R, Sullivan D, Mao C, Wang C, Sobral B, Tsujii J and Ananiadou S. Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. *BMC Bioinformatics.* 2012.

Rak R, Rowley A, Black W and Ananiadou S. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database (Oxford).* 2012: bas010.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM and Difazio SP. Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Res.* 2012; **22**(1): 95-105.

Sears KT, Ceraul SM, Gillespie JJ, Allen ED, Jr., Popov VL, Ammerman NC, Rahman MS and Azad AF. Surface proteome analysis and characterization of surface cell antigen (Sca) or autotransporter family of *Rickettsia typhi*. *PLOS Pathog.* 2012; **8**(8): e1002856.

Verhoef L, Williams KP, Kroneman A, Sobral B, van Pelt W and Koopmans M. Selection of a phylogenetically informative region of the norovirus genome for outbreak linkage. *Virus Genes.* 2012; **44**(1): 8-18.

Warren A, Setubal JC. *Using entropy estimates for DAG-Based ontologies*. ISMB SIG Bio-Ontologies. 2012.

Wattam AR, Inzana TJ, Williams KP, Mane SP, Shukla M, Almeida NF, Dickerman AW, Mason S, Moriyon I, O'Callaghan D, Whatmore AM, Sobral BW, Tiller RV, Hoffmaster AR, Frace MA, De Castro C, Molinaro A, Boyle SM, De BK and Setubal JC. Comparative genomics of early-diverging *Brucella* strains reveals a novel lipopolysaccharide biosynthesis pathway. *MBio.* 2012; **3**(5).


MIS
MEDICAL INFORMATICS SYSTEMS

SUMMARY

The Medical Informatics and Systems (MIS) Division was created in 2010, in recognition of the increasing role that medicine is playing in the lives of Americans, indeed all people in the world. Larger amounts of our GNP are consumed in medical care, and the quality of that medical care is continuously improving due to innovation. The MIS division is conducting cutting edge research in the -omics (genomics, epigenomics, transcriptomics, proteomics), genetics, systems biology and translational medicine. Our goal is to create new actionable information that will impact the work in the clinic, and the approach we have taken is to seed this new division with faculty who will bring fresh ideas, resulting in high impact research output: high impact manuscripts, new research funding, extensive collaborations and new intellectual property as a precursor to translating the findings to the clinician. VBI's newest division is well on its way, as can be seen below in the research vignettes, and in our publications and research funding successes.

GARNER LABORATORY

Microsatellite genome project

The study of families and populations with complex diseases (cancer, heart disease, neurological diseases and more) indicates that there is a significant heritable component, which intense genomics studies of single base changes (Single Nucleotide Polymorphisms, SNPs) have failed dramatically to explain. It is the Garner lab's hypothesis that another major genomic component, the understudied and difficult to measure repeated DNA (e.g., ATATATATAT) or microsatellites, can explain this missing heritability. They are understudied largely because of technological limitations, although they are known to play a role in certain diseases, and are the basis for forensics and paternity testing. Recently, the Garner lab has developed technological approaches that for the first time enable rapid, precise, inexpensive, exhaustive and simultaneous measurement of the approximately two million highly variable microsatellite loci in our genome. The lab's initial studies indicate that there are major new genomic destabilization mechanisms that globally modify microsatellites, thus potentially altering very large numbers of genes. Dr. Garner further hypothesizes that studies of microsatellites and the elucidation of these mechanisms will have a major impact on the understanding of complex diseases and traits. The laboratory's research program will greatly expand basic understanding of an understudied and underappreciated area of science by delivering: a new robust methodology; new genomic sciences knowledge that spans fundamental processes involved in speciation, evolution, complex traits and disease; rapid advances in our genomic understanding of complex disorders, starting with cancer and proceeding to neurological, cardiac and other diseases; and the foundation for development of a host of new, translatable applications that include genetic engineering for AgBio, disease predisposition and diagnostic genetic tests, pharmacogenomics tests, and drug-able biotech therapeutic leads, to name a few.

Using new array-based technological and computational approaches to study microsatellite DNA sequences, the Garner lab has made some astounding observations. For instance, all current approaches are biased and thus yield data that is inadequate for their understanding. In certain diseases (cancer and neurological diseases), and possibly universally, unique, reproducible and statistically significant changes in tens of thousands of mostly AT-rich microsatellites spread across the genome in a coordinated way. This same pattern because it is in the germline is predictive of certain cancers (breast, colon, childhood liver and more) and neurological diseases (autism and more). Unique global patterns of microsatellite variation are seen in cells which have been exposed to a carcinogen. All current genomic assemblies are deficient at the approximately two million genomic loci that contain these hypervariable repetitive motifs for which we have a combined laboratory and bioinformatic strategy to correct.

We must make an additional quantum leap in sequence production at these loci to use this new technological sophistication to efficiently and thoroughly acquire data for the two million repeated regions in the genome, thus enabling us in one fell swoop to study changes (polymorphism) at each position. This information, considered individually and en masse, will enable us to characterize the mechanism that coordinates these changes. This will simultaneously enable us to exploit them as predictors of disease, begin to understand the very role that these quantitative elements can have on our fitness to adapt to new environmental challenges, and they may also help us better understand the fine details that differentiate variation in a species from what would be considered a new species. It is imperative that this be investigated at a mechanistic level to uncover the processes that could explain why there are so many microsatellites and why they apparently sometimes act in concert.

The Garner lab's research has several components: 1) the development of robust experimental techniques to enrich genomic DNA for microsatellite regions and then acquire using next generation sequencers the accurate genotype for all regions simultaneously (the lab has recently advanced this technique and are now ready to expand our sample size); 2) the development of computational approaches to analyze this data and then leverage it against the wealth of other genomic and genetic disease data; 3) deploy 1 and 2 against a diverse and statistically relevant collection of human genomic samples that include 'disease free' individuals, a sampling of different cancers and neurological disorders, and a host of appropriate cell lines which have been subjected to different environmental stressors (chemical carcinogens, scarce and abundant nutrients, and more) for sub- and multi-generational time scales and known cell line mutants with disrupted DNA replication and error correction functions. It is expected that a significant amount of analysis of these massive datasets (billions of DNA sequences) will identify the nuances in the consistent patterns of change, the functions which alter these patterns and the correlation of these patterns with known gene information (disease and developmental relationships). The Garner lab anticipates iterating experiments with hypothesis refinements to converge upon understandings of the origins (mechanisms) of such changes and their implication in disease and evolution/development.

Microsatellites can vary continuously while SNPs have only four possible states (C,G,A,T) which may thus explain other continuously varying aspects of life (susceptibility to disease, response to therapies, physical and mental measures). This work will address a major unfulfilled component of the human genome project; that is, to enable the accurate measurement and then find the connection between our entire DNA sequence and biological systems. This will profoundly change how we use that understanding to devise ways to manipulate and exploit our DNA and that of other species to truly control our quality of life, our health and that of many other species upon which we depend through biotechnology.

Text mining: Similarity searching advances health IT/clinician decision support systems and professional ethics

The Garner lab has advanced a number of research areas by employing text mining, specifically text similarity searching to sequester relevant documents (information) and then using that information to make observations. Previously the lab used this approach to develop web-based tools to support scientific research, including eTBLAST, for reference finding, expert finding, hypothesis generation and for the study of ethics (specifically publication ethics) research which has been published in Nature, Science and other journals. This code is still provided by MIS/VBI and is widely used by journal editors and reviewers, in addition to scientists and students. Using that basic code, the Garner lab has developed two additional applications: 1) a clinician decision support system and 2) a study of grant overlap and fraud.

Health IT/clinician decision support systems

Medical records are rapidly being digitized to electronic medical records (EMRs). The EMR is formally defined as a computerized medical record created in an organization that delivers care, such as a hospital or physician's office. Although EMRs

improve administration, billing and logistics, an open research problem remains as to how doctors can leverage EMRs to enhance patient care. The concept of EMRs has been around since 1972 when it was introduced by the Regenstreif Institute. In 1991, the Institute of Medicine envisioned that by the year 2000, doctors should have adopted the usage of computers in their daily practice.

In the previous decade, the popularity of EMRs caught on and many hospitals and clinics started adopting them. The trend is expected to accelerate given the US government's push for EMR adoption by 2015 via the Health Information Technology for Economic and Clinical Health Act. The main purpose for which electronic medical records have been used is to make billing more efficient and accurate, and maintain a thorough record of patient history. It also makes ordering lab tests, tracking and recording them much more automated. However, a debate exists among physicians who do not see it benefiting them personally or helping them serve the patient better. Instead, some perceive EMRs as more of a barrier with time spent entering information into the electronic system which could be better used by spending that time with the patient. In other words, the concerns involving a return of investment or benefit to the physician are some of the main factors preventing the meaningful acceptance of an EMR management system.

There are a number of factors which make it challenging to extract information from EMRs. Each doctor has his or her own individual style of entering information and patient details, and the entries may be interspersed with acronyms-some of them being standard whereas some personalized ones have developed over the course of years. Apart from that, some of the information may be present in figures whereas others are in text. The presence of synonyms, which may be referring to the same medical conditions and the intricacies inherent to the English language as such do not make it any easier. In view of these factors, some systems have been developed in recent years to help extract information from EMRs and provide doctors with a value added system. Some of the early systems included MedLEE which converted the EMR from its natural language form to XML format, the MMTx which was developed by the NLM and was used to extract concepts from chief complaints entered in ambulatory electronic medical records. The cancer Text Information Extraction System (caTIES), on the other hand, is used to extract information from free text surgical pathological reports which helps locate resources related to tissues.

This report describes a system that analyzes a patient's evolving EMR in context with available biomedical knowledge and the accumulated experience recorded in various text sources including the EMRs of other patients.

The aim of the Clinician Decision Support (CDS) Dashboard is to provide interactive, automated, actionable EMR text-mining tools that help improve both the patient and clinical care staff experience. The CDS Dashboard, in a secure network, helps physicians find de-identified electronic medical records similar to their patient's medical record, thereby aiding them in diagnosis, treatment, prognosis and outcomes. Of particular value in cases involving complex disorders, it also allows physicians to explore relevant medical literature, recent research findings, clinical trials and medical cases.

A pilot study done with medical students at the Virginia Tech Carilion School of Medicine showed that 77% of them found the CDS Dashboard to be useful in aiding patient care for doctors and 88% of them found it useful for aiding medical students pedagogically and 88% of them found the tool user friendly. The Garner lab's aim is to develop a tool which can be generic yet customizable, modular and extensible.

The CDS Dashboard uses a multidisciplinary approach to the construction of the system, including computer science, medicine, biomedical research, and human-machine interfacing and is proven to be of potential value to clinicians and medical practitioners **(Fig. 1)**.

The CDS system is designed to run concurrently with existing EMR software and provide doctors with pertinent information with respect to the

current patient's EMR they are viewing. The pertinent information includes de-identified EMRs of other patients who have faced similar medical problems, relevant medical literature, case reports and clinical trials. The CDS system has been designed such that doctors can customize what patient information should be displayed, from which sources they want relevant information drawn and in which order. One of the main advantages of the CDS is that it allows the doctors control over how they want to use the software **(Fig. 2)**.

Professional ethics

In February and March of 2012, funded grant summaries (2,262,605 records corresponding to 858,717 grant/contract summaries in total) were downloaded from public web sites in the US for the NIH, NSF, DoE, DoD and Susan G. Komen for the Cure®. The total number of grant/contract summaries which could be compared were reduced by 227,380 (26.5%) because they contained less than 50 words, and thus could not be accurately processed by the lab's computational methods. Thus 2×10^{11} $((858{,}717\text{-}227{,}380)^2 / 2)$ text similarity comparisons were possible, reducing the overall sensitivity to 54%. The number and years of grant summaries obtained varied greatly across agencies, with awards summing to over $200B (NIH 1985-2012, NSF 1985-2012, DoE 1995-2009, DoD 1993-2011, and Komen 2003-2011).

Each grant summary was compared to all others, and thus intra- as well as inter-agency funding of projects could be evaluated.

Here, the Garner lab focused exclusively upon evaluating inter-agency funded grant similarity using software to calculate similarity scores for all possible grant pairs. The text similarity engine, eTBLAST, calculates a similarity score between each pair of grant summaries following the same approach we established to identify potential plagiarized scientific literature. This approach relies on a two pass full-text algorithm to accurately and efficiently calculate similarity scores based on weighted keywords and sentence alignments. Using this score the pairwise

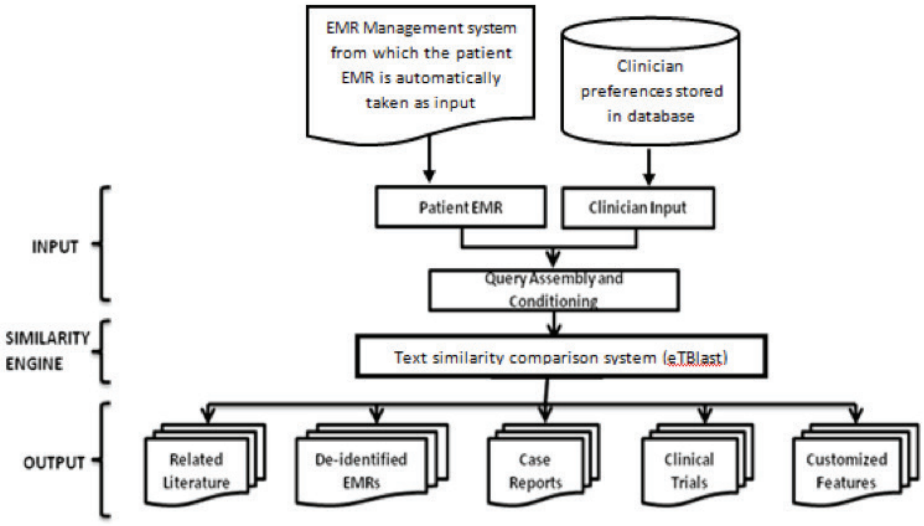


Figure 1. Flow chart of CDS Dashboard. The input query is formed by automatically reading the electronic medical record from the EMR Management System, eg: EPIC and combining them with the user/clinician preferences stored in the database. These preferences can be changed at any point of time by the user. The EMR is then preprocessed and fed as input into the text similarity comparison system (eTBlast), which then compares it to various other indexed databases and displays the results according to user preferences.

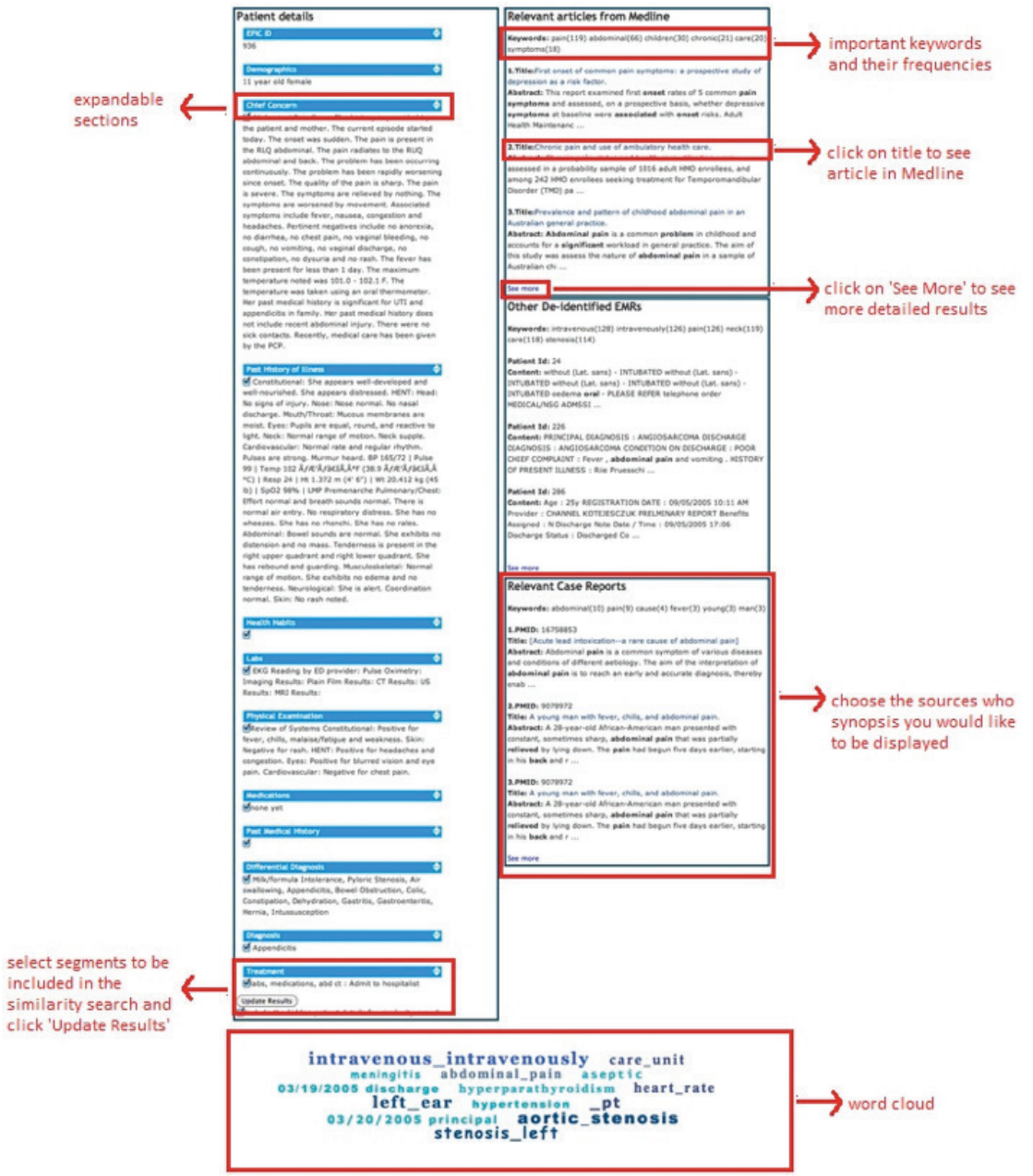


Figure 2: Typical user interface customized according to user preferences. The user interface contains the patient's EMR automatically fetched from the EMR Management System and displayed according to user preferences. The results from similarity search are displayed on the right hand side and the word cloud is displayed below.

similarity score was used to rank them in descending amounts of similarity to prioritize for manual review. For this collection of documents, the similarity score for all pairs ranged from 0 to 1.8, and all documents were reviewed down to a score of 0.8, an arbitrary cutoff which provided a sufficient large sampling. The focus was narrowed to only the 1573 grant pairs of summaries out of the 2×10^{11} possibilities having the highest similarity score. Manual analysis found 167 pairs of summaries that had been suspected overlapping of proposed work.

Among the most similar summaries with identical or highly similar specific aims, objectives, goals and hypotheses, the Garner laboratory identified the following trends:

1. Similarity among science projects, where funding may overlap in time or be sequential. Grants may be funded simultaneously or be recycled later to another agency. Approximately 31% of the grants/contracts ran concurrently.
2. Similarity between a science project and support for (sometimes multiple) post doctoral fellows or dissertation support of the science project PI/mentor.
3. The later grant proposes studies cited as preliminary data in the earlier grant of the pair.
4. Similar grants with different PIs (same or different institutions) funded by different agencies.
5. Substitution of different gene names, drugs or organ sites that follow otherwise similar research plans.
6. Some of the pairs of similar funded projects in the biomedical domain were searched for associated literature in PubMed, and occurrences were found where only one agency was acknowledged. Some occurrences revealed additional highly similar summaries (not included in our totals).
7. Potential overlap is a problem today. In the most recent 5 years (2007-2011), there were 39 grant pairs involving over $20M in the later grants identified.

This work and previous work on professional plagiarism may also indicate a need for a new emphasis area on scholarly writing ethics (manuscripts, grant applications, web, etc.). This area would have some unique and differential characteristics, including quantitative and cross disciplinary analysis. This could provide for publications and grantsmanship what NIH/NHGRI's ELSI (Ethical, Legal and Social Issues) program provided for genomics.

Heterogeneous high-performance computing for the life and medical sciences

Advances in High-performance Computing in the MIS division has taken several forms, including projects funded by the NSF, nVidia Foundation and startup funds to Dr. Garner. These funds have enabled us to advance the analysis of genomic and medical informatics big data, especially in genomics and in text mining. The preliminary data for these projects came from the creation of Shadowfax, a heterogeneous supercomputer with microprocessor, GPU and FPGA processing power, very large storage and data bandwidth, and other features.

Shadowfax was defined by Dr. Garner as a collaborative machine, with initial funding from his startup package, and now has grown significantly via contributions from others at VBI. This machine has enabled many to complete major scientific studies and with the aid of our partners, the VBI IT team was able to make the machine immediately productive, with a compute utilization of greater than 85%. This machine has enabled the analysis of thousands of genomes to discover new cancer biomarkers and was used to analyze almost one million grant summaries in an ethics study of possible grant double-dipping. Shadowfax is being used by many other groups at VBI and the campus community of Virginia Tech. One computational component of Shadowfax includes hybrid-core machines from Convey Computing. (For more information on Shadowfax, see the Computing Core section). Outside of US government agencies, VBI has the largest installation of Convey hardware. This long running collaboration ultimately enabled the MIS division to win an NSF cyberinfrastructure grant. This NSF grant supports the development of "personalities" or unique programs written for the Convey machines that combine gate array processing and standard microprocessors, the development of new on-the-fly compression

and decompression approaches to address the storage and bandwidth challenges that are typical of biomedical computing big data problems, the demonstration of these components via a series of biomedical applications and a summer school to educate and train 16 researchers from diverse institutions. These efforts represent a highly collaborative project involving VBI, electrical and computer engineering and computer science faculty. Lastly, Dr. David Mittelman is supported by the NVIDIA Foundation, a philanthropic arm of nVidia Corporation, the company that manufactures all graphic processor units (GPUs). In this project, his group develops and demonstrates a biomedical toolbox that enables users to quickly develop applications on GPUs, especially for genomic problems. These three programs have proven to be very successful, establishing a highly collaborative, cross-campus and beyond HPC environment, and providing the foundation for future more ambitious HPC installations at Virginia Tech.

MITTELMAN LABORATORY

Building genomic analysis pipelines

Genome sequencing technologies promise to revolutionize our understanding of genetics, evolution, and disease by making it feasible to survey a broad spectrum of sequence variation on a population scale. However, this potential can only be realized to the extent that methods for extracting and interpreting distinct forms of variation can be established. The error profiles and read length limitations of early versions of next-generation sequencing technologies rendered them ineffective for some sequence variant types, particularly microsatellites and other tandem repeats, and fostered the general misconception that such variants are inherently inaccessible to these platforms. At the same time, tandem repeats have emerged as important sources of functional variation. Tandem repeats are often located in and around genes, and frequent mutations in their lengths exert quantitative effects on gene function and phenotype, rapidly degrading linkage disequilibrium between markers and traits. Sensitive identification of these variants in large-scale next-gen sequencing efforts will enable more comprehensive association studies capable of revealing previously invisible associations.

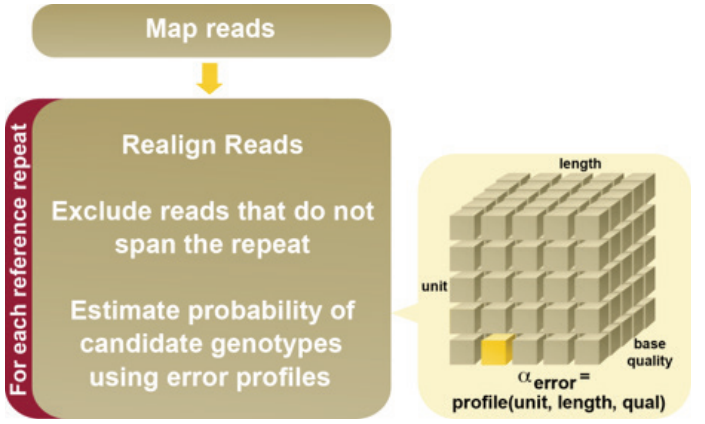


Figure 1. Reads are mapped and realigned, and a set of reads spanning reference repeats are retained. Genotypes are assigned with consideration of the a *priori* error rate, which comes from the appropriate error profile and is used in the prior distribution of allele and error probabilities. The probability of each genotype suggested by the data is estimated in a Bayesian fashion and the most probable genotype among these is called.

The Drosophila Genetic Reference Panel

The Mittelman lab is performing population-scale analyses of microsatellite repeats and other mutation types using whole-genome data from almost 200 inbred isolates from the Drosophila Genetics Reference Panel (DGRP), a collection of extensively phenotypically characterized isolates from a single natural population, to uncover processes underlying repeat mutation and to enable associations with behavioral, morphological and life-history traits. The DGRP lines are each derived from single female founders of a natural fly population, and bred to near-isogeny by 20 generations of full-sibling matings. Therefore, although alleles may differ among lines, in the absence of mapping, alignment or sequence errors, all reads from a single inbred line mapped to a specific microsatellite locus should support the same allele. The assumption of homozygosity permits the use of internal concordance among the various reads within each inbred line to assess the relative accuracy of alternative approaches and tune heuristics for mutation detection.

RepeatSeq: Accurate microsatellite genomes from HTS data

Repetitive sequences are biologically and clinically important because they can influence traits and disease, but repeats are challenging to analyze using short-read sequencing technology. The Mittelman lab has developed a tool for genotyping microsatellite repeats called RepeatSeq, which uses Bayesian model selection guided by an empirically derived error model that incorporates sequence and read properties. To validate this approach, Dr. Mittelman applied RepeatSeq to high-coverage genomes from the 1000 Genomes Project and evaluated performance and accuracy against other methods. The software uses common formats such as VCF for compatibility with existing genome analysis pipelines **(Fig. 1)**. The lab currently has the most accurate published method for repeat mutation detection.

All datasets and genomics software are available and fully supported via GitHub and SEQanswers, the largest online genomics community.

Publications:

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM,…Mittelman D, and Gibbs RA. The Drosophila Genetic Reference Panel: A community resource for analysis of population genomics and quantitative traits. *Nature*. Feb 9, 2012;482:173-8.

Fondon J, III, Martin A, Richards S, Gibbs RA, and Mittelman D. Analysis of microsatellite variation in *Drosophila melanogaster* with population-scale genome sequencing. *PLOS ONE*. 2012; 7(3): e33036.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, and Mittelman D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res*. Oct 22, 2012.

Li J, Schmieder R, Ward RM, Delenick J, Olivares EC, and Mittelman D. SEQanswers: An open access community for collaboratively decoding genomes. *Bioinformatics*. 2012; 28(9):1272-3.

XIE LABORATORY

Toward the understanding of epigenomic dynamics

The primary research interest of the Xie lab is to understand the molecular mechanisms underlying epigenetic transitions during important biological processes associated with human complex diseases. Toward this goal, Dr. Xie emphasizes the development of high-throughput sequencing approaches for data generation and the implementation of computational tools for -omics data analysis. In particular, the lab is interested in strategies to monitor the fidelity of DNA methylation inheritance, assess methylation variation within and between cell populations, identify true epimutations, and, eventually, discover novel therapies. Our previous studies include bio-technique development to facilitate targeted epigenome sequencing, CpG islands and Alu repeats in particular; and model establishment to assess DNA methylation variations and simulate epigenetic state transitions during stem cell differentiation, reprogramming and disease genesis. We have successfully applied techniques and bioinformatics tools to identify diagnosis markers for diseases.

Development of targeted bisulfite sequencing technology

DNA methylation plays a crucial role in the control of gene expression, genomic imprinting and the maintenance of genome integrity. It primarily occurs at CpG dinucleotides, which are not uniformly distributed in vertebrate genomes. Some genomic regions, such as CpG islands and Alu repeats with a high GC content and a high frequency of CpG dinucleotides, harbor many transcriptional factor binding sites. The Xie lab developed convenient and cost-effective approaches to selectively target these important genomic regions **(Fig. 1)** to uncover the methylation aberrations involved in the genesis and progression of human diseases, including cancer.

Design and implementation of computational tools for -omics data mining

The challenges in an epigenomic study are not limited to data generation but also data interpretation. Traditionally, DNA methylation data analysis is based on the determination of the average methylation level of one or more contiguous CpG sites. Such a conventional way cannot dissect DNA methylation patterns, which are herein defined as the combination of methylation statuses of contiguous CpG dinucleotides in a DNA strand. Since most DNA methylation datasets were derived from a cell population rather than a single cell, they comprise

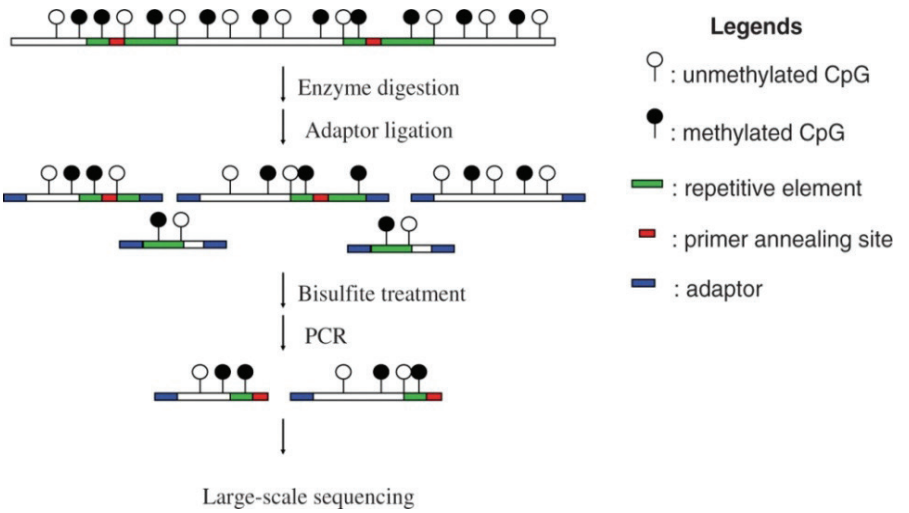


Figure 1: Alu-anchored genome-wide bisulfite sequencing

the information of epigenetic variation within a cell population. To better decode epigenetic data, we introduced the concept of "methylation entropy" to dissect the variation of DNA methylation patterns in a given cell population. Such analysis enables us to assess epigenomic heterogeneity and distinguish the methylation changes resulting from deterministic or stochastic events. We anticipate that the introduction of methylation entropy will bring a new dimension to DNA methylation data analysis.

Other publications

Hoxha E, Lambers E, Xie H, et al. Histone deacetylase 1 deficiency impairs differentiation and electrophysiological properties of cardiomyocytes derived from induced pluripotent cells. *Stem Cells*. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22915496

Sittig LJ, Herzing LB, Xie H, Batra KK, Shukla PK, Redei EE. Excess folate during adolescence suppresses thyroid function with permanent deficits in motivation and spatial memory. *Genes Brain Behav*. 2012;11:193–200. http://www.ncbi.nlm.nih.gov/pubmed/22050771

Wang D, Liu X, Zhou Y, et al. Individual variation and longitudinal pattern of genome-wide DNA methylation from birth to the first two years of life. *Epigenetics*. 2012;7:594–605. http://www.ncbi.nlm.nih.gov/pubmed/225229l0

Wang M, Xie H, Shrestha S, Sredni S, Morgan GA, Pachman LM. Methylation alterations of WT1 and homeobox genes in inflamed muscle biopsies from untreated juvenile dermatomyositis suggests self-renewal capacity. *Arthritis Rheum*. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22674142

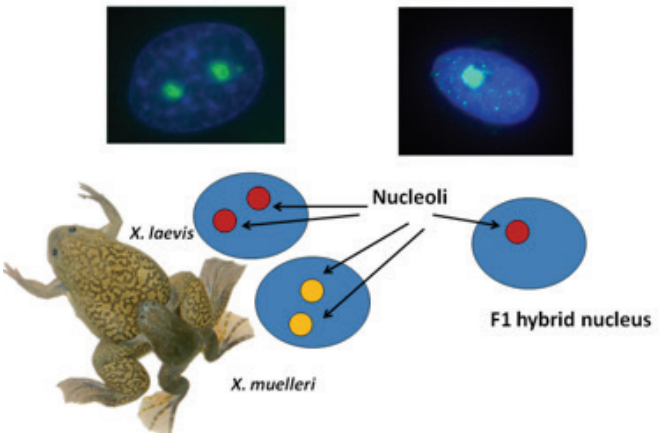
Zhang X, Wallace AD, Du P, et al. DNA methylation alterations in response to pesticide exposure *in vitro*. *Environ Mol Mutagen*. 2012;53:542–549. http://www.ncbi.nlm.nih.gov/pubmed/22847954

Zhang X, Wallace AD, Du P, et al. Genome-wide study of DNA methylation alterations in response to diazinon exposure *in vitro*. *Environ Toxicol Pharmacol*. 2012. http://www.ncbi.nlm.nih.gov/pubmed/22964155

MICHALAK LABORATORY

The RNA world meets disease and evolution - ribosomal DNA instability

The research program in the Michalak lab focuses on epigenetics and functional genomics of reproductive defects and stress resistance. The lab employs an integrative approach with a variety of molecular and computational techniques, including next-gen sequencing of entire genomes and transcriptomes, to rigorous understanding of how complexity elaborates from genomes through gene regulatory networks, and how it evolves over generations in response to the environment. Research in the Michalak lab has implications for cancer, chromosomal abnormalities, birth defects, and stress responses. For brevity, one project is highlighted below.



Ribosomal DNA and the logic of genomic conflicts

Despite decades of research on ribosomal gene silencing, little is known about how rRNA genes are selected for inactivation at the genome-wide level. With their evolutionary dynamics seemingly unconstrained by Mendelian laws of inheritance, clusters of ribosomal DNA provide a fascinating paradigm of genome instability and plasticity with fundamental consequences for cellular metabolism. Dr. Michalak has recently discovered that rDNA in *Xenopus* hybrids is expressed predominantly from maternal copies, thus proving that parental genome imprinting can be well established in organisms without intimate parent-offspring interactions during embryo development. Unexpectedly, he has also found that the pattern of maternal rRNA expression is independent of nucleolar dominance (decrease in the number of nuclei and species-specific effects). Species-origin-dependent rDNA dominance was observed in the first generation of backcrossing, but not in F1 or further backcrosses, showing that nucleolar dominance is even more complex and puzzling than previously thought. The mechanistic determinants of rDNA imprinting and nucleolar dominance are currently being pursued.

Ribosomal DNA (rDNA) has considerable autonomy within the genome, potentially facilitating intragenomic conflicts, with certain rRNA gene copies transmitted from one generation to the next in a biased manner. There are a number of rDNA peculiarities consistent with DNA selfish element dynamics. For example, excessive and highly variable numbers of identical or nearly identical copies, many of which remain transcriptionally silent, are spread around the genome, segregate later than other parts of the genome, can extrachromosomally circularize, and invade other selfish elements such as B chromosomes and germline-limited DNA. As a result of genetic incompatibilities between evolutionarily divergent genomes, potential genomic

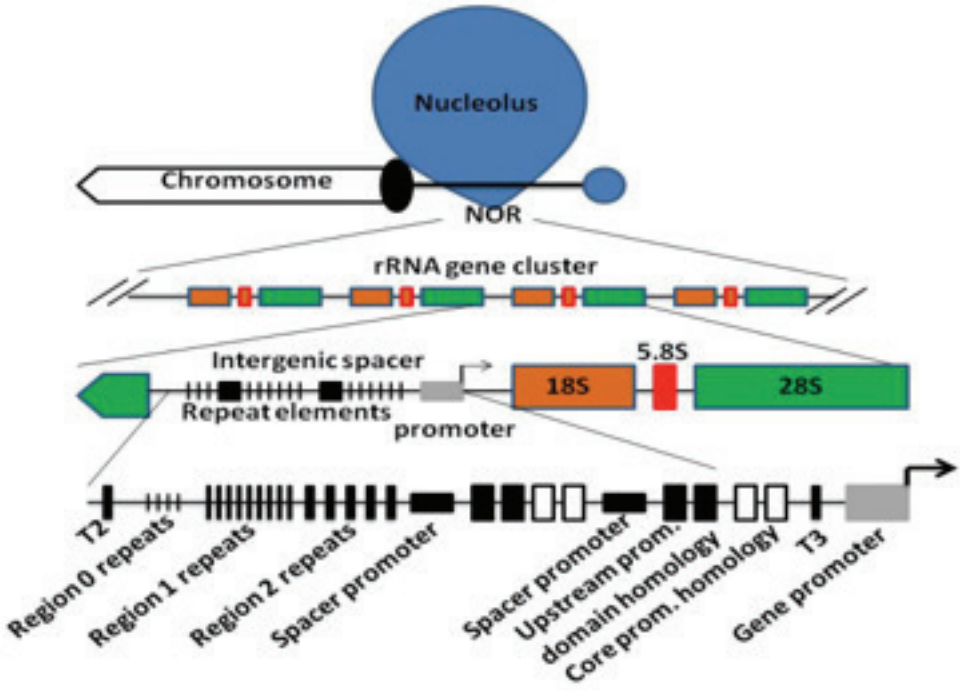


Figure. 1: A) A pictorial representation and fluorescent images of X. laevis and F1 hybrid cells displaying nucleolar dominance. In the fluorescent images the nucleoli are in green, immunostained with X. *laevis* nucleolar protein while the nucleus is stained with DAPI. **B)** NOR of X. *laevis*. An annotated close-up of the X. *laevis* enhancer region is displayed on the lowest portion.

conflicts and "arms races" among rDNAs should be most pronounced in interspecies hybrids and manifest as either origin-dependent expression imprinting (violations of Mendel's first law) or transmission ratio distortion (violations of Mendel's second law). Indeed, the former often results in nucleolar dominance, whereby hybrids consistently suppress formation of nucleoli and expression of rRNA from one of the parental species **(Fig. 1)**, whereas the latter can be observed as a unilateral rRNA gene loss in hybrid species. The Michalak lab hypothesizes that these two seemingly disparate parent-origin effects, one a property of somatic cells and the other a property of germline, are signatures of a single evolutionary process operating on differences in the chromatin structure and epigenetic instructions, engaging selfish rDNA, and creating genomic conflicts. Epigenetic differences between homologous chromosomes, presumably maintained for proper pairing and recombination, may also result in parental-origin-dependent monoallelic expression and non-random segregation.

Publications

Koroma AP, Jones R, and Michalak P. Snapshot of DNA methylation changes associated with hybridization in *Xenopus*. *Physiol Genomics*. 2011; 43(22): 1276-80.

PECCOUD LABORATORY

Genetic Design Automation: A new framework to elucidate structure-function relations in genomic sequences

Synthetic biology aims to make genetic systems more amenable to engineering, which has naturally led to the development of Computer-Aided Design (CAD) tools. Experimentalists still primarily rely on project-specific ad-hoc workflows instead of domain-specific tools, suggesting that CAD tools are lagging behind the front line of the field. The Peccoud group is focused on overcoming the scientific hurdles that have limited the productivity gains anticipated from existing tools. Our goal is to demonstrate that the immediate value of efforts to develop CAD tools is the formalization of genetic design rules that determine the complex relationships between genotype and phenotype in natural and engineered genomes. We specifically focus on understanding the genomic implementation of the cell cycle regulatory network in the yeast genome and the engineering of RNA viruses as vectors for developing vaccines against emerging pathogens.

CAD tools for synthetic biology

Several groups have been developing CAD solutions for synthetic biology, yet the transcriptional complexity of published artificial gene networks has been leveling off since 2005. After ten years of high expectations and hype in synthetic biology, engineering biological systems has proved more challenging than anticipated. The lack of sufficient tools in synthetic biology has spurred intense efforts to develop CAD software. Unfortunately, experimental synthetic biologists still rely largely on project-specific, ad-hoc development processes that combine construct assembly, data collection, data analysis, and mathematical modeling.

Five recent reviews have comprehensively covered the current state of computational tools for synthetic biology. CAD tools for synthetic biology facilitate the design of larger systems from smaller genetic parts by providing users with visual, textual, or programming language-like interfaces, or automatically generate designs from intended function. These tools assume that data such as sequence and description are attached to each part by the user or in some database. The aggregated parts sequences can then be leveraged to produce the corresponding physical DNA. Many tools include some level of functional modeling capabilities, which rely on the user providing the necessary equations and/or parameters **(Fig. 1)**.

Thus far in synthetic biology, relatively simple design goals such as "exhibits oscillations" have advanced to increasingly sophisticated goals such as "fast, robust, tunable oscillators" or "synchronized oscillators". As the field moves towards real-world applications, tools that can adequately predict functionality from design will be indispensible. Similarly, designers need to begin considering alternative design approaches and corresponding comparison metrics to move beyond proof of concept designs. Recently announced design-to-spec competitions like CAGEN (Competitive Assessment of Genetically Engineered Networks) and GenoCon (International Rational Genome Design Contest) aim to help address this need.

The Peccoud group focuses on overcoming the scientific hurdles that have limited the productivity gains anticipated from existing CAD tools. It also aims to demonstrate that the scientific value of these efforts is not in promised productivity gains, but in the formalization of genetic design rules. Formalizations inherently test commonly held conceptions of how genetic systems work and consequently drive investigation into one of the most fundamental questions of genetics: How does phenotype arise from complicated networks of elements coded in the genotype?

The slow maturation of EDA

CAD tools are ubiquitous in nearly all fields of engineering. They provide two primary functions: simplifying common tasks and making designs convenient for communication and evaluation. For example, blueprints are designed much faster on a computer than with traditional drafting. Moreover, CAD tools can generate layouts of each floor, 3D renderings of the building or models of the building's thermal performance.

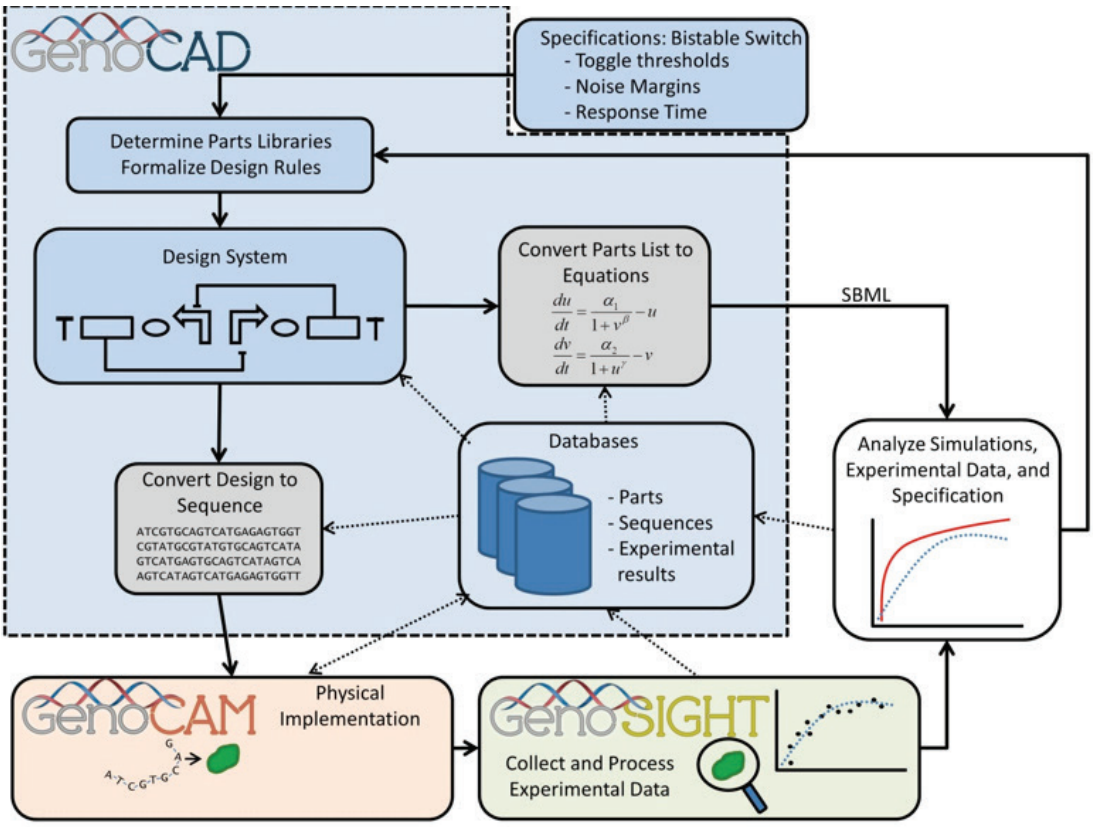


Figure 1: **The GenoFAB tool chain supports GDA design flow.** Synthetic biology projects typically rely on iterative workflows composed of different tasks. The GenoFAB tool chain is composed of GenoCAD, a design tool, GenoCAM, a manufacturing tool, and GenoSIGHT, a testing tool. These software applications support different phases of the project workflow. For instance, the development of a genetic switch will start by expressing the design objective as a list of quantitative requirements: input toggle thresholds, noise margins, switching response time, etc. Once the objective is specified, it is possible to develop a list of genetic parts useable for the project. The choice of biological parts will involve factors such as use of the parts in prior projects, quality of the data characterizing the parts function, or intellectual property considerations. The formalization of design rules often takes place in parallel to the parts library development. Design rules may express rules such as whether it is acceptable to have polycistronic expression cassettes or if the design should be split between different plasmids. Only after parts have been selected and a strategy has been agreed upon is it possible to start designing constructs. In the fabrication phase, the construct is assembled usually by combining de novo gene synthesis and cloning of existing DNA sequences. Users use molecular biology software suites to facilitate assembly or order the sequence from a gene synthesis company. Experimentalists insert the synthetic DNA molecule into the host of choice and collect phenotypic data. Experimental data is then processed, for example by reducing microscopy images into time series of quantitative data. Performance is evaluated by considering simulations, experimental data, and the original specifications.

In electronics, the development of consistent suites of CAD tools is called Electronic Design Automation (EDA). EDA utilizes iterative "design flows," where key design processes are looped back on themselves until the design meets required specifications. Abstract representations of an electronic design in EDA can be organized into a hierarchy consisting of high-level description, logical description, and physical layout.

The spectacular success of EDA over the last 50 years provides an inspiring model for synthetic biology. The synthetic biology counterpart of EDA is sometimes referred to as Bio-Design Automation. However, the alternative name Genetic Design Automation (GDA) may better emphasize that synthetic biology focuses more on engineering DNA molecules than other biological objects. Many have proposed that synthetic biologists leverage expertise in the design of electrical circuits, and in the same way GDA can draw from the development of EDA. In EDA, Hardware Description Languages (HDLs) are a special category of programming languages used to formally describe immensely complicated designs in a compact way. These languages rely extensively on the existing abstraction that digital circuits operate under the laws of Boolean algebra, which allows hugely complex circuits to be designed reliably.

This assumption does not exist for GDA. Exploratory work on an HDL for GDA has progressed under the assumption that similar enabling assumptions will emerge. In parallel, efforts within the EDA community to describe analog and mixed analog-digital circuits with HDLs are ongoing. The challenges of extending HDLs to analog circuits are very similar to the challenges faced in GDA, and indeed some works have explored these similarities.

The next sections describe three difficult problems that the Peccoud group is working on to make the level of automation in EDA achievable with GDA. There is mounting evidence that the first generation of GDA tools will not be able to ignore some of the most challenging problems currently faced by the EDA community.

The scientific gaps facing GDA

Transitions between DNA sequence, model, and fabrication are currently hindered not so much by the implementation of CAD tools, but rather by three difficult scientific challenges: (i) predictability of components, (ii) decoupling of design and fabrication and (iii) experimental characterization methods.

GenoCAD: Off-the-shelf components

Since 2009, the Peccoud group has been developing GenoCAD, a web-based CAD tool for synthetic biology available at www.genocad.org. This effort is funded by NSF Award 0850100.

One of the popular visions for synthetic biology describes catalogues of clearly defined genetic parts that can be easily combined into larger genetic constructs with predictable biological function. This vision motivated the development of the BioBrick assembly standard and the Registry of Standardized Genetic Parts, a database of BioBrick compatible parts. Tools that aggregate models of basic genetic components to form system-scale models are being developed, but the lack of data sheets listing quantitative parameters characterizing the parts behavior has hampered the use of these tools for designing artificial gene networks. Projects such as BioFAB are attempting to address this issue by characterizing large numbers of parts and standardizing data collection techniques.

Recent efforts to quantitatively characterize the effects of different parts on gene expression is revealing a complex landscape of context-dependencies that somewhat challenges the assumption that parts can be characterized in isolation. For instance, the ribosomal binding site (RBS) sequence was first assumed by many in the field to determine translation efficiency in prokaryotes independently from the downstream coding sequence. However, sequences around the translation start site can influence the secondary structure of the mRNA, which has long been known to play a crucial role in the translation rate. Tools utilizing thermodynamic models are now available to predict the translation initiation efficiency in prokaryotes using sequences both upstream and

downstream of the translational start site. Coupling between translation and transcription elongation rates also represents a challenge to the standardization of components, though the assumption that initiation, not elongation, is the rate limiting step in transcription may be a valid approximation. As a result, tools that can predict behavior based on sequence, thermodynamics or other methods are emerging as increasingly attractive.

The above issues can be avoided by characterizing on a gene-by-gene or device-by-device level, a trend already apparent in the field. Creation of device variants or automatically generated devices should consider the many context dependencies that affect parts. Yet, even such low levels of granularity might prove to have unexpected context dependencies. Computational studies, inspired by impedance-matching in electronics, have demonstrated an effect termed "retroactivity" in which the performance of one genetic device is influenced by connecting a downstream device. Just as electronic circuit designers are currently running into major power limitations, synthetic biologists are almost certain to run into limits on the many ingredients necessary for gene expression. How the availability of resources within a cell impacts the performance of individual genetic components and devices will also become an important consideration.

GenoCAM: Recoupling of design and fabrication

In 2012, the Peccoud group received a grant from NSF (Award 1241328) to adapt methods from industrial systems engineering to the optimization of DNA fabrication workflows. This award has been used to lay the foundations of a Computer Assisted Manufacturing system for DNA fabrication.

Historically, recombinant DNA technologies were so limited that fabrication constrained design to the point that software focused almost entirely on assisting cloning rather than design of function. The recent emergence of generic DNA fabrication methods, including standardized assembly of genetic parts or *de novo* gene synthesis led to the emergence of DNA-sequence design as a new scientific problem. Because it is now possible to assume that generic DNA fabrication processes can assemble any sequence genetic engineers can imagine, design and fabrication tend to be considered orthogonal engineering problems.

In EDA, the assumption of the Boolean abstraction has allowed fabrication to be considered mostly orthogonal to the rest of the design process. As circuit densities have rapidly increased, fabrication constraints have become more closely intertwined with other constraints such as timing delay and power consumption. Consequently, increasingly integrated tools consider constraints from multiple design domains simultaneously. Genetic design is still in the process of moving away from fabrication technologies that constrain the design space to achieve complete independence between design and manufacture. For instance, BioBrick assembly standards have moved from the original standard precluding assembly of fusion proteins to proposed standards allowing fusions or scarless assembly. Most recently, single-step assembly methods have become popular.

Despite the simplification decoupling offers, there are distinct advantages to recoupling design and manufacturing during the design phase. Poor design strategies can create manufacturing problems. For example, repeated use of the same parts can cause sequence verification difficulty and structural instabilities. Sequences with high GC content are notoriously difficult to amplify. Even though experienced gene synthesis companies will be able to synthesize most DNA sequences ordered by their customers, price and time to delivery vary greatly with sequence complexity. Ignoring such manufacturing constraints during the design phase will significantly increase the cost and duration of GDA cycles. Since many projects require the characterization of large numbers of design variants, the cost and time to fabricate these designs is still one of the bottlenecks of the GDA loop. Sophisticated design strategies are necessary to formalize manufacturing constraints and optimize fabrication without serious detriment to function. For example, tools that can adjust codon bias or match non-unique sequences to function could ameliorate manufacturing concerns. Sample tracking tools could suggest reuse of sequence

segments resulting in shortened and cheaper assembly cycles. Algorithms to optimize fabrication processes will have to be connected to design tools to give designers an immediate appreciation of the manufacturing cost of candidate designs.

GenoSIGHT: Parts Characterization

The Peccoud group has been developing GenoSIGHT, an imaging cytometry platform to characterize the noisy dynamics of gene networks at the single cell level. This line of research is supported by NIH through two grants focused on modeling the regulatory network controlling cell division in yeast (awards R01-GM078989 and R01-GM095955).

The problem of part definitions extends to the ways experimental data are collected, used and shared. Models vary from project to project and typically have only one or two fluorescent reporters as measured outputs. Since the models typically have many more parameters to estimate, finding parameter sets that are predictive from the many possible sets that match a given dataset is difficult. Combined with the range of measurement techniques used in the lab and the unknown impact of even small changes in experimental conditions, standardization and reuse of collected data are challenging.

Though not necessarily an explicit step in standard EDA design flows, Design for Test (DFT) is a background constraint throughout the process. As electronic circuits have become more complex, validation testing has become correspondingly more difficult. First, immense complexity makes exhaustive validation intractable, so intelligent minimization of test programs has evolved along with sophisticated testing equipment. Second, the density of modern chips has made accessing nodes on internal layers without unintended performance effects a major problem. In living cells, the problem is much worse. Unknown cellular mechanisms, genetic instabilities, molecular noise, measurement accuracy, and inability to measure key components have prevented the verification of synthetic biology designs.

DFT in GDA starts by ensuring consistency between the design specification, experimental characterization methods, and mathematical paradigms used to model the behavior. For example, models based on deterministic equations can be supported by cell culture assays, whereas stochastic models call for single cell observations. Experimental design, common in some other engineering fields and recently emerging as a hot topic in systems biology, can guide experimentalists to optimal sets of measurements and avoid non-informative data collection. Depending on the system, this approach can also predict data sets that will determine all model parameters to within a given tolerance. Similar techniques are frequently used in EDA to design intelligent testing programs, particularly for System-on-a-Chip designs where analog components commingle with digital circuits. Not considering the experimental characterization during the design phase may lead to designs that simply cannot be adequately tested.

The problem of collecting time-course data from inside living cells has been transformed by the emergence of fluorescent proteins. However, just like making an internal node in a chip accessible for test affects its performance, the maturation rate and protein half-life buffer fluorescence signals from the physiological events of interest. Unfortunately, even parameters related to measurements such as the *in vivo* maturation rate of fluorescent proteins are difficult to determine accurately due to cell-to-cell variation and unknown dependencies on host strain, metabolic state, and environmental conditions. Also, the use of genes tagged with fluorescent domains is common, but the effects poorly understood. For example, how does a fluorescent tag affect the degradation rate of the fusion protein? Experimental work is needed to answer these questions.

New measurement devices like GenoSIGHT facilitate the ability to determine *in vivo* parameter values and their dependencies on the continuously variable cellular environment. The emergence of time lapse fluorescent microscopy has allowed scientists to measure the dynamics of molecular mechanisms in individual cells. GenoSIGHT integrates optical equipment, image processing software, and microfluidic systems. Noise analysis connects the processed data to models of the underlying

mechanisms. Current image processing and analysis are typically post-processing steps, but GenoSIGHT represents a new generation of instruments that can acquire raw images, process them in real time, and implement data reduction algorithms to extract high-level statistics for comparison with design-phase models **(Fig. 2)**. Microfluidic systems allow for complicated input control of environmental parameters that should open avenues to applying some advanced testing techniques seen in EDA DFT, such as frequency response analysis.

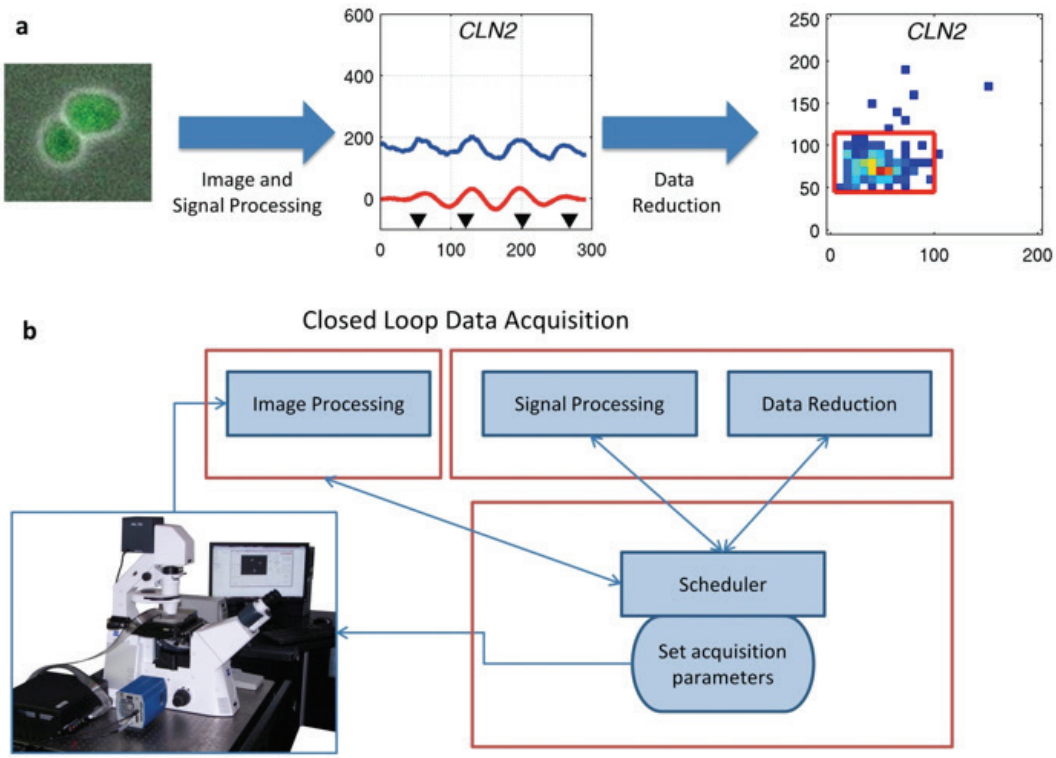


Figure 2: **GenoSIGHT overall architecture. (a)** Using time-lapse microscopy for characterizing the dynamics of gene networks requires the development of custom suite of image and signal processing software along with data reduction algorithms. The mathematical models used to reduce movies into high-level statistics are necessarily related to the models used to design the gene network as ultimately experimental data need to be reconciled with model predictions. **(b)** Microscopy movies have traditionally been analyzed in a post-processing step. GenoSIGHT performs data analysis in real time by the computer controlling the microscope and the microfluidic system giving the user an experience similar to the use of a flow-cytometer. This information can be used by the user to manually interact with the cells under observation. GenoSIGHT also includes algorithm to program the instrument to take specific actions such as changing the growth medium in response to specific behaviors of the cell populations.

A scientific renewal

It is the Peccoud lab's belief that synthetic biology will lead to a better understanding of biology. In this spirit, the immediate value of GDA does not lie in its potential to accelerate the progress of experimental synthetic biologists. In the short term, the focus should be on demonstrating that efforts to develop GDA tools are better justified as attempts to formalize genetic design principles. The assumption that models used in engineering can be extrapolated to biology can easily and rightfully be challenged by biologists. A more effective way of using GDA to engage the dialogue between engineers and life scientists might be to present GDA models as formal and compact representations of biological hypotheses. GDA then becomes a framework to express and test biological hypotheses, a form of scientific investigation common in the life sciences.

It is important to consider here two fundamental differences between EDA and GDA. The dynamics of genetic networks are largely determined by the interactions between large macromolecules confined to the small volume of a living cell. As a result, the dynamics of a genetic network are inherently stochastic in nature. There is even mounting evidence that many regulatory processes are based on molecular noise instead of merely attempting to mitigate its negative effects. Electronic circuits use so many electrons that they behave deterministically. However, a consequence of the miniaturization and increasing power efficiency of silicon devices is a drastic reduction of the fluxes of electrons and a concomitant increase of the intrinsic electronic noise. It will be interesting to see if and how the EDA and GDA communities will work together to solve the problems associated with design automation of noisy systems.

Another important difference between GDA and EDA is that complexity in EDA was derived, while complexity in GDA was evolved. EDA has progressed by structured, rational improvements on the mathematical formalisms that express physical realities, incrementally allowing higher and higher complexity. As a result, the emergence of high-level function can always be traced to the lowest level components. On the other hand, since genetic systems evolved by random mutation, it is not clear that they follow rigorous design rules, and we cannot yet trace high-level function back to the low-level components in most cases. Synthetic biologists are walking in the footsteps of 50 years of effort by molecular geneticists to understand the design rules of genetic systems. Yet, the engineering mindset provides a new spark. The understanding of a genetic mechanism is truly put to the test when an engineer attempts to use the general principle to build something new.

Formalization of these principles tests the theory and opens new areas of investigation when the theory is found lacking. A prime example of this is the aforementioned RBS calculator. Attempts to use "standard" ribosome binding sites failed and led to predictive thermodynamic models. The possibility to deoptimize the sequence of viral genes by taking advantage of codon pair bias is opening new research directions to better understand translation. At a higher level of organization, the refactoring of the T7 genome resulted in reduced fitness that it is not completely understood.

In the current state, synthetic biology remains painfully slow, prohibitively expensive and excessively labor-intensive. As an example, consider the progression from two 2002 theoretical papers on genetic oscillators to corresponding experimental publications in 2008 and 2010. The unifying vision of a seamless GDA flow provides a collaborative framework for large interdisciplinary teams rather than relying on exceptional individual investigators familiar with all aspects of the design process. There is still a great deal of foundational work and biological discovery remaining before GDA materializes into suites of software tools facilitating design to specification entirely *in silico*. In the short term, closing some of the capability gaps will catalyze the emergence of more integrated teams that better handle the complex interdependencies between design, fabrication, and measurement. The tools these teams will generate may not have the elegance of an integrated solution, but they will provide new computational resources that should

percolate beyond the confines of the synthetic biology community to benefit a larger population of life scientists.

Publications 2012

Lux MW, Bramlett BW, Ball DA, Peccoud J. Genetic design automation: Engineering fantasy or scientific renewal? *Trends Biotechnol.* 2012; 30(2): 120-126.

Marchand JA, Peccoud J. Building block synthesis using the polymerase chain assembly method. *Methods in Molecular Biology.* Clifton, NJ. 2012; 852: 3-10.

Peccoud J, Isalan M. The PLoS ONE Synthetic Biology Collection: Six Years and Counting. *PLOS* ONE 2012; 7(8): 7.

Wilson ML, Cai Y, Hanlon R, Taylor S, Chevreux B, Setubal JC, Tyler BM, Peccoud J. Sequence verification of synthetic DNA by assembly of sequencing reads. *Nucleic Acids Res.* 2012.

CORE LABORATORIES

GENOMICS RESEARCH LABORATORY (GRL)

With the advent of modern next-generation DNA sequencers and bioinformatics software analysis tools, we are beginning to realize the great potential of genomics. What once took over 10 years and $3 billion to accomplish – the sequencing of the first human genome – can now be accomplished for less than $3000 in under one month, and shortly will be under $1000 in one day. The impact of this technological revolution is just starting to be realized in a significant way and is leading to the development of therapeutic treatments, diagnostic tests, new forms of genetically modified foods, and countless other products and treatments. As human genome sequencing becomes even less costly and faster, it is easy to imagine sequencing as a common procedure performed during routine physical examinations, much like cholesterol or blood pressure measurements are used today. However, sequencing genomes is, in many regards, the easy part. Interrupting and managing the enormous amounts of data produced by next-generation sequencing is the difficult part of the equation. To handle both the generation and analysis of genomic sequences, VBI has created a new laboratory, the Genomics Research Laboratory (GRL) and staffed it with experienced technicians and PhD-level bioinformatician scientists.

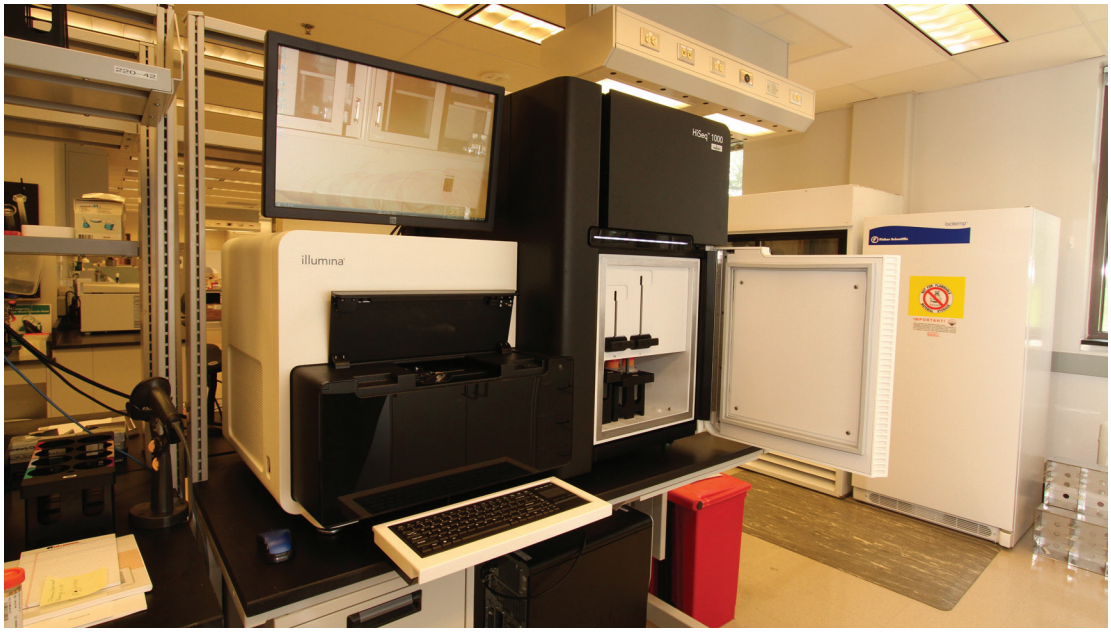
Over the past year, the Genomics Research Laboratory (GRL) has undergone a paradigmatic transformation. This can be seen not only in the name change from Core Laboratory Facility (CLF) to GRL and the upgrade of the next-generation sequencing equipment, but also in the shift from simply sequencing genomics to studying the function and interaction of genes. To help facilitate this change in focus, we hired a new Laboratory Manager, Dr. Saikumar Karyala. Dr. Karyala came to us from the University of Cincinnati with a rich background in functional genomics and next-generation sequencing equipment. He provided whole-genome analysis services at the University of Cincinnati Genome Sequencing and Microarray Core, usually working with the university and Children's Hospital. He is an experienced researcher in the fields of breast and prostate cancer.

Dr. Karyala will help the Genomics Research Laboratory achieve its aim of being a "one-stop shop," providing researchers access to cutting-edge equipment and experienced scientists. A one-stop shop means that the GRL users are provided with expert consultation on experimental design, access to production-quality laboratories, and a scientific staff to assist in the development of new technical protocols and procedures.

Housed within the GRL is VBI's Data Analysis Core (DAC), a dedicated multi-user resource that provides state-of-the-art data analysis support for biologically related research. The DAC was created to remove data analysis as an obstacle to performing the highly complex, data- and computing-intensive experiments required to solve modern problems in life science research. The Data Analysis Core offers a range of services that result in manuscript-ready figures and text. Working with the researcher and the Genomics Research Laboratory, the DAC scientists are available to provide guidance and support in designing and executing on properly statistically powered and controlled experiments in the most cost effective manner.

Beyond experimental design, the DAC offers many services including: statistical expression analysis (microarray or RNAseq), functional analysis, genome assembly, variation analysis, gene prediction and annotation. While the DAC operates in a cost-recovery setting, to fully realize its benefits, the DAC scientists must work closely with GRL users to interpret data in the biological setting for which it was collected. The DAC works with data from next-generation sequencing and many other platforms and multiple experimental formats (CHIPseq, RNAseq, DNA-reseq, SNP-seq, etc).

With access to the computing centers at VBI, the GRL and DAC can take on large collaborations that would not be possible at many sequencing cores, such as the analysis of gene activation in maturing domesticated turkeys. This is a large collaboration involving Drs. Rami Dalloul of VT and Kent Reed of the University of Minnesota. The goal of this work



is to examine the changes in gene expression that occur in various organs of the domestic turkey as the bird grows and matures. "The scope of this project is really quite immense and computationally challenging. Not only are we faced with the millions of sequence reads that result from RNA-seq approaches, we are dealing with a non-model genome that is as of yet incompletely sequenced. These challenges require a more hands-on approach and complex solutions," said Dr. Kent Reed, professor of veterinary and biomedical sciences at University of Minnesota. To fully enable this project, the DAC had to rethink how to accomplish large computationally intensive projects. Here, researchers in the GRL brought to bear the full power of VBI's Computing Core, leveraging more than 1000 cores of state-of-the art computers and using more than 20 TB of disk space.

Currently, GRL scientists are working with Virginia Tech researchers to analyze samples from a variety of sources, including human, turkey, peanut, poplar tree, and more. These analyses have resulted in the identification of potential biomarkers for cancer, and the first complete assembly of the genome for the domestic turkey, to mention two accomplishments. Other current projects range from the study of blight-resistant legumes to discovering the culprit behind widespread illness amongst buffalo. GRL members have been involved in several grant submissions this year, and have co-authored several publications in leading scientific journals.

In an effort to more fully support the Virginia Tech Life Sciences Initiative, the GRL pioneered a small grant program to benefit Virginia Tech researchers. The genetics of the yellow fever mosquito and poison ivy urushiol genes are among the 16 research topics funded through Virginia Bioinformatics Institute and Fralin Life Sciences Institute's small grants program. VT researchers from eight departments and institutes won awards ranging from $7500 to $19,000 to conduct research in the life sciences using the specialized next-generation sequencing equipment, facilities, and services offered by the Genomic Research Laboratory at the Virginia Bioinformatics Institute. Vying for a portion of a $200,000 award pool, fifty-two research teams from VT submitted proposals totaling $700,569. Drs. Stanley Hefta, director of the Genomics Research Laboratory at VBI, Boris Vinatzer, professor, of plant pathology, physiology and weed science at VT and Liwu Li, professor of inflammation biology and immunology at VT, evaluated the proposals based on the scientific merits of the research and the likelihood that additional extramural funding and publications would result from the study.

INFORMATION TECHNOLOGY AND COMPUTING CORE

Information Technology (IT), including the Computing Core (CC), at VBI is responsible for providing a secure, stable, and manageable infrastructure supporting data-intensive research. In addition, the CC provides a robust computational environment for the institute's administration. The staff is composed of multiple systems administrators, database administrators, programmers, and help desk technicians.

VBI and Virginia Tech have undertaken a large investment in providing advanced research computing for our scientists and researchers. Our goal has been to provide superior compute and data storage infrastructure specifically to meet the needs of the life and medical sciences. The investment alone over the last two years at VBI has totaled more than $3 million. The combination of VBI's state-of-the-art laboratory facilities and instrumentation, along with the scale of its compute facilities, makes us one of the premier computational biomedicine institutes in the country.

Computational resources encompass more than 3000 processor cores of various architectures distributed across 220 servers and several clusters, with over 10 TBs of memory, and more than 1.5 PBs of disk storage. Current facilities include three data centers totaling 3000 square feet and a Situation Awareness/ Training room. Two large UPS (Uninterruptible Power Systems) with matching diesel generators provide backup power for all of the computational systems. The first data center is utilized by administrative computing; the second houses our research and HPC systems; while the third is home to our tape library.

VBI was the first building on VT campus to install GbE (1 Gigabit Ethernet) throughout the entire facility, including desktops. Additionally, in the data centers there are two dedicated storage area networks: 8Gbps Fibre Channel and 10Gbps Ethernet.. Data centers are connected via dual GbE to the campus backbone, which in turn is connected to Network Virginia and Internet2 via redundant 10 GbE links. In preparation for our expanding storage needs and the necessity of exchanging data with collaborators, VBI is in the process of implementing an internal 10GbE network and a dedicated 10GbE connection from VBI data centers to the campus network.

HPC storage resources include the following:

- Primary home filesystem space consists of 315TB residing on SGI IS16000 system. Data is accessible via GPFS (General Parallel File System) through a dedicated QDR (Quad Data Rate) InfiniBand link at rates of over 2 GB/s (GigaBytes per second). With additional disk drives and disk enclosures this storage will peak at over 10 GB/sec. (Please note: This system will be used for home directories for NDSSL for use with the DIIME proposed system. The NDSSL directories can be allocated on a separated file system within this storage).
- Fast scratch directories for HPC also take advantage of a second SGI IS16000 storage system, which provides 315TB of space via a QDR InfiniBand link.
- 270TB of Dell Compellent enterprise SAN storage for scientific databases and other critical file systems. This storage system provides the Oracle database space used by NDSSL.
- 320TB of Pillar Axiom600 are available via SAN (Storage Area Network) for additional redundancy on critical file systems such as scientific databases.
- Archive storage complimented by over 1.25PBs of tape pool, seamlessly integrated by SGI's DMF (Data Migration Facility). The system is attached via 10GbE network.
- A StorageTek SL8500 tape library using 5TB T10000-C tapes with total capacity of 10,000 tapes provides tape backup and archive; peak capacity would be well over 50 PB.

The extensive networking and storage capability is utilized by VBI's high-performance computing (HPC) resources for life science and medical applications. Large data sets, both of internal and external origin, are stored and processed by applications running on clusters and servers.

Shadowfax

VBI's Shadowfax cluster consists of 140 dual-socket nodes, two quad-socket large memory nodes, and eight special purpose FPGA (Field Programmable Gate Array) nodes. It's currently an all-purpose general VBI resource computational system used for sequence alignment, graph manipulation and other embarrassingly parallel jobs. This system provides approximately 630TB of high-performance GPFS storage for both scratch and home directories.

Shadowfax is a hybrid cluster system featuring standard compute nodes of Intel Westmere and Sandy-bridge processors. The system also includes two large memory nodes with 40 cores and 1 TB of memory each. There are also eight nodes with FPGA that provide accelerated genomics processing. The system also features NVIDIA Tesla and Kepler GPGPUs.

This system has been operational for over two years. Shadowfax has undergone three expansions over the last year, leading to a factor of three increase in the number of nodes and performance. The current system is heavily utilized and must be continually upgraded to keep up with NDSSL's vast compute needs.

Expertise

VBI Core Computational Facility staff has been at the forefront of computational technology for life and medical sciences. Over the past decade system administrators have deployed multiple HPC clusters and enterprise systems, including Sun Microsystem's

SunFire 15000, IBM eServer Cluster 1300 and IBM p655 based cluster. Storage products from various vendors span multiple protocols and network infrastructures.

CC staff are proficient and cross-trained in all areas of UNIX and Linux system administration, high-performance computing, high availability systems and enterprise storage management. They also maintain a vast library of scientific software and accompanying knowledge base, along with providing training to end-users. A team of helpdesk computer system specialists aids with the task of day-to-day issue management and customer support. Database administrators provide a key role in design and implementation of world-class databases housing projects for NIH, NSF and other sponsored programs.

All personnel participate in regular training, both internal and external, taking advantage of classes taught at Virginia Tech and professional organizations. For example, every year the system administrators attend a weeklong SANS Institute workshop. Each CC member also takes advantage of University Organizational and Professional Development (UOPD) branch at Virginia Tech. Annual conferences, seminars and workshops include SuperComputing, HPC User Forum and HPC Bootcamp.

Kevin Shinpaugh, Director of the Computing Core and IT at VBI, has worked in the area of high-performance computing for over nine years and in information technology for over 25 years. As Director of Research Computing at Virginia Tech, he managed multiple projects, including building and managing System X, which was the third fastest supercomputer in the world November 2003. This project also included the upgrade of the Virginia Tech data center, which was done in a record time of five weeks.

In addition to the dedicated HPC systems administrators, the CC team is composed of two other Linux/UNIX system administrators, five help desk computer technicians, four developers and two database administrators. The expertise and responsibility areas of the CC staff cover all aspects of administrative and research computing at the institute.

Collectively, the VBI team of molecular biologists, scientist, data analysts, IT specialists, and project mangers has a vast experience in dealing with a multitude of compute and data intensive life and medical science applications and the associated hardware required to run them.

In addition, IT/CC delivers the following services:

- Server (dedicated and shared/virtual) hosting
- High-performance computing
- Storage and backup
- Database hosting
- Building and supporting unique information technology systems
- Application development, maintenance, and deployment
- Data support
- Enabling services to support research collaborations


GCI
GRANTS + CONTRACT INFORMATION

COMPLETED PROJECTS

Project Title: Accelerated biomethanation of sequestered carbon dioxide and paraffin in coal beds-Phase II
PI: Andrew Scott
Co-PIs: Biswarup Mukhopadhyay
Source of Support: Altuda Energy Corporation (DOE)
Total Amount Awarded: $339,000
Total Award Period Covered: 08/07/06-08/09/12

Project Title: Alternaria and proteases in innate and adaptive Th2 immunity
PI: H. Kita
Co-PIs: Christopher Lawrence
Source of Support: NIH/NIAID
Total Amount Awarded: $1,864,935
(VBI Portion: $252,463)
Total Award Period Covered: 12/01/07-11/30/12

Project Title: Characterization of mobile RNAs in a host-parasite interaction
PI: James Westwood
VBI PI: Genomics Research Lab
Source of Support: NSF
Total Amount Awarded: $44,567
Total Award Period Covered: 09/01/09-08/31/12

Project Title: Collaborative research: Estimating speciation/reticulation boundaries in asexual *Alternaria:* A genomics approach
PI: Christopher Lawrence
Source of Support: NSF
Total Amount Awarded: $267,559
Total Award Period Covered: 07/01/09-06/30/12

Project Title: Collaborative research: modeling interaction between individual behavior, social networks and public policy to support public health epidemiology
PI: Madhav Marathe
Co-PIs: Christopher Barrett, Stephen Eubank, Richard Beckman, Anil Vullikanti
Source of Support: NSF HSD
Total Amount Awarded: $750,000
(VBI Portion: $540,000)
Total Award Period Covered: 07/01/07-09/30/12

Project Title: Collaborative research: NECO: A market-driven approach to dynamic spectrum sharing
PI: Samir R. Das
Co-PIs: Anil Vullikanti, Achla Marathe, Madhav Marathe, Himanshu Gupta, Milind M. Buddhikot
Source of Support: NSF
Total Amount Awarded: $1,311,969
(VBI Portion: $490,000)
Total Award Period Covered: 09/01/08-08/31/12

Project Title: Comprehensive cancer center of Wake Forest University, biostatistics consulting
PI: Ina Hoeschele
Source of Support: WFUHS (NIH/NCI)
Total Amount Awarded: $223,055
Total Award Period Covered: 02/01/06-01/31/12

Project Title: Compute the cure
PI: David Mittelman, Wu Feng
Source of Support: NVIDIA Foundation
Total Amount Awarded: $100,000
(VBI Portion: $50,000)
Total Award Period Covered: 05/01/11-04/30/12

Project Title: COPASI: Biochemical network modeling and simulation software
PI: Pedro Mendes
Source of Support: NIH
Total Amount Awarded: $2,057,649
(VBI Portion: $1,656,878)
Total Award Period Covered: 09/01/07-08/31/12

Project Title: Development of a versatile mutant resource in tomato for functional genomics analysis
PI: Andy Pereira
Co-PIs: Richard Veilleux
Source of Support: USDA/NRI
Total Amount Awarded: $399,500
(VBI Portion: $248,671)
Total Award Period Covered: 08/01/07-07/31/12

Project Title: Expedited high-affinity ligand generation from biomarker sequence by integration of data minig with high-throughput screening
PI: Harold Garner
Source of Support: Lynntech, Inc.
Total Amount Awarded: $51,238
Total Award Period Covered: 02/01/10-10/31/12

Project Title: Genome-wide functional analysis of *Phytophthora* small RNAs
PI: Nikolaus Grunwald
Co-PIs: Brett Tyler, J. Carrington
Source of Support: USDA-CSREES Functional Genomics of Agriculturally Important Organisms
Total Amount Awarded: $980,000
(VBI Portion: $286,185)
Total Award Period Covered: 04/01/08-01/14/12

Project Title: High-performance computing methods for inference state assessment and course of action analysis in large socio-technical methods
PI: Christopher Barrett
Co-PIs: Richard Beckman, Henning Mortveit, Madhav Marathe
Source of Support: DTRA
Total Amount Awarded: $1,425,000
(VBI Portion: $1,200,000)
Total Award Period Covered: 03/11/09-07/14/12

Project Title: How do oomycete and fungal effectors enter host plant cells?
PI: Brett Tyler
Co-PIs: Christopher Lawrence, Daniel Capelluto, Shiv Kale
Source of Support: NSF
Total Amount Awarded: $500,000
(VBI Portion: $327,217)
Total Award Period Covered: 09/01/09-08/31/12

Project Title: IDBR: Development of CytoIQ, a 5D adaptive cytometer to measure the noise of dynamics of molecular interactions in individual live cells
PI: Jean Peccoud
Source of Support: NSF
Total Amount Awarded: $160,000
Total Award Period Covered: 06/15/10-05/31/12

Project Title: Intertwined evolutionary history of methanogenesis and sulfate reduction: An analysis of putative sulfate reduction pathway genes in methanogens
PI: Biswarup Mukhopadhyay
Source of Support: NASA
Total Amount Awarded: $146,014
Total Award Period Covered: 08/27/09-08/26/12

Project Title: Lawrence Livermore Short Course
PI: Madhav Marathe
Source of Support: Lawrence Livermore National Laboratory
Total Amount Awarded: $74,424
Total Award Period Covered: 12/01/10-10/31/11

Project Title: Molecular genetic network of water use efficiency in rice
PI: Andy Pereira
Co-PIs: Niranjan Baisakh, Hei Leung
Source of Support: USDA
Total Amount Awarded: $349,958
Total Award Period Covered: 09/01/09-08/31/12

Project Title: Nutritional regulation of human Crohn's Disease by conjugated linoleic acid
PI: Josep Bassaganya-Riera
Source of Support: Cognis Deutschland Gmbh & Co.
Total Amount Awarded: $153,681
Total Award Period Covered: 09/01/05-12/31/11

Project Title: Obesity-related inflammation
PI: Josep Bassaganya-Riera
Source of Support: European Union
Total Amount Awarded: 100,000 Euros
Total Award Period Covered: 01/16/08-01/15/12

Project Title: REU: IDBR: Development of CytoIQ, a 5D adaptive cytometer to measure the noise of dynamics of molecular interactions in individual live cells
PI: Jean Peccoud
Source of Support: NSF REU
Total Amount Awarded: $16,373
Total Award Period Covered: 05/01/11-05/31/12

Project Title: The pathogenesis of chronic rhinosinusitis
PI: H. Kita
Co-PIs: Christopher Lawrence
Source of Support: NIH/NIAID
Total Amount Awarded: $2,448,735
(VBI Portion: $689,585)
Total Award Period Covered: 05/01/07-04/30/12

Project Title: TRMS: Cereal drought stress response and resistance networks
PI: Andy Periera/Kristy Collins
Co-PIs: Oswald Crasta, Ruth Grene, Niranjan Baisakh, Yinong Yang
Source of Support: NSF
Total Amount Awarded: $2,400,000
Total Award Period Covered: 09/01/09-08/31/12

CURRENT PROJECTS

Project Title: A systems approach to iron metabolism in cancer cells
PI: Reinhard Laubenbacher
Source of Support: NIH
Total Amount Awarded: $383,566
Total Award Period Covered: 08/01/11-07/31/13

Project Title: ABI development: Representation, visualization, and modeling of signaling pathways in higher plants
PI: Lenwood Heath
Co-PIs: Ruth Grene, Andy Periera
Source of Support: NSF
Total Amount Awarded: $1,057,336
(VBI Portion: $320,882)
Total Award Period Covered: 05/01/11-04/30/15

Project Title: Bioinformatics Resource Centers Portal
PI: Bruno Sobral
Source of Support: NIH
Total Amount Awarded: $5,633,226
Total Award Period Covered: 10/01/09-09/30/14

Project Title: BREAD: Engineering novel resistance against fungal and oomycete pathogens in developing country crop plants
PI: Brett Tyler
Co-PIs: S. Xiao, M. Guiltinan
Source of Support: NSF BREAD
Total Amount Awarded: $1,450,975
(VBI Portion: $698,249)
Total Award Period Covered: 04/15/10-03/31/13

Project Title: CAREER: Cross-layer optimization in cognitive radio networks in the physical interference model based on SINR constraints: Algorithmic foundations
PI: Anil Vullikanti
Source of Support: NSF
Total Amount Awarded: $450,000
Total Award Period Covered: 02/01/09-01/31/14

Project Title: CC-NIE network infrastructure: ASCED - an advanced scientific collaboration environment and DMZ
PI: Mark Gardner
VBI PI: Keith Bisset
Source of Support: NSF
Total Amount Awarded: $296,581
(VBI Portion: $20,198)
Total Award Period Covered: 01/01/13-06/30/14

Project Title: Coenzyme F420-dependent enzymes in mycobacteria
PI: Biswarup Mukhopadhyay
Source of Support: NIH
Total Amount Awarded: $435,648
Total Award Period Covered: 06/13/12-05/31/14

Project Title: Collaborative Research: ABI Innovation: PlantSimLab: A simulation laboratory for plant biology
PI: Reinhard Laubenbacher
Co-PIs: Stefan Hoops, John McDowell, Brett Tyler
Source of Support: NSF
Total Amount Awarded: $881,510
(VBI Portion: $799,077)
Total Award Period Covered: 06/15/12-05/31/15

Project Title: Collaborative Research: Coupled models of diffusion and individual behavior over extremely large social networks
PI: Madhav Marathe
Co-PIs: Keith Bisset, Xizhou Feng, Dimitris Nikolopoulos
Source of Support: NSF
Total Amount Awarded: $1,450,000
(VBI Portion: $819,942)
Total Award Period Covered: 09/01/09-08/31/13

Project Title: Collaborative Research: Simulation of contagion on very large social networks with Blue Waters
PI: Keith Bisset
Co-PIs: Xizhou Feng
Source of Support: NSF
Total Amount Awarded: $18,999
Total Award Period Covered: 05/01/09-04/30/13

Project Title: Comprehensive National Incident Management System (CNIMS)
PI: Christopher Barrett
Source of Support: DTRA
Total Amount Awarded: $10,012,079
(VBI Portion: $9,746,456)
Total Award Period Covered: 10/01/11-09/30/14

Project Title: Computational biomathematics: Toward optimal control of complex biological systems
PI: Reinhard Laubenbacher
Source of Support: ARO
Total Amount Awarded: $414,932
Total Award Period Covered: 07/01/09-09/14/13

Project Title: COPASI software for modeling and simulation of biochemical networks
PI: Pedro Mendes
Source of Support: NIH
Total Amount Awarded: $3,412,261
Total Award Period Covered: 09/01/07-05/31/16

Project Title: Determining feasibility and scalability of a life/medical science hybrid-core based platform
PI: Harold Garner
Co-PIs: Peter Athanas, Kristy Collins, Liqing Zhang
Source of Support: NSF
Total Amount Awarded: $1,309,316
(VBI Portion: $1,264,559)
Total Award Period Covered: 08/01/11-07/31/14

Project Title: Diffusion on complex networks: Algorithmic foundations
PI: Anil Vullikanti
Source of Support: DOE
Total Amount Awarded: $750,000
Total Award Period Covered: 03/01/10-02/28/15

Project Title: Digging for metabolic fossils in *Methanocaldococcus jannaschii*
PI: Robert White
Co-PIs: Biswarup Mukhopadhyay
Source of Support: NSF
Total Amount Awarded: $1,054,866
(VBI Portion: $213,628)
Total Award Period Covered: 01/01/12-12/31/15

Project Title: Disease chronicity in juvenile dermatomyositis (JDM): Epigenetic clues
PI: Hehuang Xie
Source of Support: Children's Hospital of Chicago
Total Amount Awarded: $42,458
Total Award Period Covered: 08/01/12-07/31/13

Project Title: Dissecting functional cooperation among subunits in a catalytic ribonucleoprotein
PI: Venkat Gopalan
Co-PIs: Biswarup Mukhopadhyay
Source of Support: Ohio State University (NSF)
Total Amount Awarded: $27,993
Total Award Period Covered: 02/15/09-01/31/13

Project Title: Drug-resistant hospital acquired infections – tools for prevention and treatment
PI: Biswarup Mukhopadhyay
Source of Support: Carilion
Total Amount Awarded: $13,715
Total Award Period Covered: 03/01/11-02/11/13

Project Title: DynSyst Special Topics: Polynomial dynamical systems over finite fields: from structure to dynamics
PI: Reinhard Laubenbacher
Source of Support: NSF
Total Amount Awarded: $277,935
Total Award Period Covered: 09/01/09-08/31/13

Project Title: EAGER: Redox regulation of methanogenesis: A basic microbial process at the interface of climate change and renewable energy REU supplement
PI: Biswarup Mukhopadhyay
Source of Support: NSF
Total Amount Awarded: $6,750
Total Award Period Covered: 05/15/12-05/31/13

Project Title: EMBERS: Early model based event detection using surrogates
PI: Narendran Ramakrishnan
Co-PIs: Achla Marathe, Anil Vullikanti, Scotlan Leman, Chang-Tien Lu, Michael Roan, Aravind Srinivasan, Tanzeem Choudhury, John Brownstein, Dipak Gupta, Johan Bollen, Kristen Summers
Source of Support: IARPA
Total Amount Awarded: $14,494,495
(VBI Portion: $2,018,526)
Total Award Period Covered: 04/01/12-03/31/15

Project Title: Epigenetic capacitors of adaptations to thermal stress
PI: Pawel Michalak
Co-PIs: Abraham Korol, Eugenia Rashkovetsky
Source of Support: US-Israel Binational Science Foundation
Total Amount Awarded: $135,000
(VBI Portion: $73,144)
Total Award Period Covered: 10/01/12-09/30/15

Project Title: Epigenome-wide association study of DNA methylation and artherosclerosis
PI: Ina Hoeschele
Source of Support: Wake Forest University (NIH)
Total Amount Awarded: $559,641
Total Award Period Covered: 07/01/09-06/30/14

Project Title: Exhaustive analysis of microsatellite loci in the 1000 Genomes Project
PI: Harold Garner
Source of Support: NIH
Total Amount Awarded: $540,000
Total Award Period Covered: 06/26/10-04/30/13

Project Title: Exploring the roles of adaptive plasticity and invasibility of alewife (*Alosa pseudoharengus*) to better manage fisheries in the Great Lakes
PI: Pawel Michalak
Source of Support: Illinois Natural History Survey
Total Amount Awarded: $89,303
Total Award Period Covered: 09/15/11-12/31/12

Project Title: Functional interactomics: Determining the roles played by members of the poplar biomass protein-protein interactome
PI: Eric Beers
Co-PIs: Allan Dickerman
Source of Support: DOE/USDA
Total Amount Awarded: $2,693,869
(VBI Portion: $392,931)
Total Award Period Covered: 08/01/07-07/31/14

Project Title: FY 2010 energy efficient building systems regional innovation cluster initiative
PI: John Burns
Source of Support: Penn State (DOE)
Total Amount Awarded: $4,999,999 (VBI Portion: $298,106)
Total Award Period Covered: 09/01/10-08/31/15

Project Title: Gene ontology terms and automated annotation for energy-related microbial genomes
PI: Biswarup Mukhopadhyay
Co-PIs: Brett Tyler, T.M. Murali
Source of Support: DOE
Total Amount Awarded: $1,142,453
(VBI Portion: $1,127,606)
Total Award Period Covered: 07/01/10-06/30/13

Project Title: Genetic determinants of visceral adiposity: A Genome-wide Association Study
PI: Ina Hoeschele
Source of Support: Wake Forest University (NIH)
Total Amount Awarded: $100,136
Total Award Period Covered: 02/01/11-01/31/13

Project Title: Genetic reassessment after induction in advanced lung cancer
PI: Michael Boyd
Source of Support: CHRB
Total Amount Awarded: $82,664
(VBI Portion: $78,216)
Total Award Period Covered: 07/01/12-06/30/14

Project Title: GenoCAD: A Computer Assisted Design environment for synthetic biology
PI: Jean Peccoud
Co-PIs: M. Czar, S. Faulkner
Source of Support: NSF
Total Amount Awarded: $1,421,725
(VBI Portion: $1,221,160)
Total Award Period Covered: 09/01/09-08/31/13

Project Title: Genomic and proteomic architecture of atherosclerosis
PI: Rebecca Will
Source of Support: Wake Forest University (NIH)
Total Amount Awarded: $3,381,302
(VBI Portion: $2,667,921)
Total Award Period Covered: 07/01/12-06/30/17

Project Title: Haplotyping and QTL mapping in pedigrees with missing data
PI: Ina Hoeschele
Source of Support: Virginia Commonwealth University (NIH)
Total Amount Awarded: $42,635
Total Award Period Covered: 06/01/07-05/31/13

Project Title: Highly multivariate quantitative trait loci mapping in systems genetics
PI: Ina Hoeschele
Source of Support: NIH
Total Amount Awarded: $760,720
Total Award Period Covered: 02/01/10-11/30/13

Project Title: ICES Large Collaborative Research: The role of space, time, and information in controlling epidemics
PI: Anil Vullikanti
Co-PIs: Achla Marathe
Source of Support: NSF
Total Amount Awarded: $295,000
Total Award Period Covered: 07/01/12-06/30/15

Project Title: INSPIRE: Modeling and optimization of DNA manufacturing processes
PI: Jean Peccoud
Co-PIs: Jaime Camelio, Kimberly Ellis
Source of Support: NSF
Total Amount Awarded: $999,531
(VBI Portion: $457,511)
Total Award Period Covered: 08/01/12-07/31/15

Project Title: Integrated management of oomycete diseases of soybean and other crop plants
PI: Brett Tyler
Co-PIs: J. Gordon Arbuckle, Jaime Blair, Carl Bradley, Thomas Chase, Martin Chilvers, Paul Esker, Dean Fraga, Loren Giesler, Doug Jardine, Nicholas Kalaitzandonakes, Reinhard Laubenbacher, Christopher Lawrence, Donald Lee, Dean Malvik, Sam Markell Saghai Maroof, John McDowell, Bill Morgan, Paul Morris, Berlin Nelson, Manuel Ospina-Giraldo, Wayne Parrott, Vipa Phuntunmart, Alison Robertson, Joao Setubal, Karen Sirum, Aurelian Tartar, Kiersten Wise
Source of Support: USDA AFRI
Total Amount Awarded: $9,281,250
(VBI Portion: $8,700,777)
Total Award Period Covered: 04/01/11-03/31/16

Project Title: Integrating top-down and bottom-up models in systems biology with applications to cell cycle control in budding yeast
PI: T.M. Murali
VBI PI: Jean Peccoud
Source of Support: NIH
Total Amount Awarded: $2,090,926
(VBI Portion: $475,023)
Total Award Period Covered: 01/01/11-12/31/15

Project Title: LLNL Fellowship
PI: Madhav Marathe
Source of Support: Lawrence Livermore National Laboratory
Total Amount Awarded: $64,357
Total Award Period Covered: 06/13/11-12/31/12

Project Title: Longitudinal study investigating the progression and pathogenesis of atypical hyperadrenocorticism in Scottish Terriers
PI: Kurt Zimmerman
Co-PI: Ina Hoeschle
Source of Support: AKC Canine Health Foundation
Total Amount Awarded: $66,226
(VBI Portion: $25,200)
Total Award Period Covered: 07/01/11-06/30/13

Project Title: Middle Atlantic Regional Center for Excellence (MARCE) for biodefense and emerging infectious disease
PI: Mike Levine
Source of Support: University of Maryland (NIH)
Total Amount Awarded: $386,625
Total Award Period Covered: 03/01/09-02/28/14

Project Title: Modeling Immunity to Enteric Pathogens (MIEP)
PI: Josep Bassaganya-Riera
Source of Support: NIH
Total Amount Awarded: $10,571,877
(VBI Portion: $8,176,696)
Total Award Period Covered: 09/15/10-09/14/15

Project Title: Modeling of infectious disease high school internship
PI: Kristy Collins
Source of Support: RTI International
Total Amount Awarded: $19,999
Total Award Period Covered: 09/01/12-08/31/13

Project Title: NetSE Large Collaborative Research: Contagion in large socio-communication networks
PI: Madhav Marathe
Co-PIs: Anil Vullikanti, Stephen Eubank, Chris Barrett
Source of Support: NSF
Total Amount Awarded: $2,998,238
(VBI Portion: $1,544,999)
Total Award Period Covered: 08/15/10-07/31/15

Project Title: Outbreak of *Acinetobacter baumannii:* Not an old friend, but a new enemy in the hospital setting
PI: Jayasimha Rao
Co-PI: Biswarup Mukhopadhyay
Source of Support: Carilion Medical Center
Total Amount Awarded: $6,950
Total Award Period Covered: 08/02/10-12/31/12

Project Title: PAThosystems Resource Integration Center (PATRIC) – Bioinformatics Resource Center (BRC) for bacterial diseases
PI: Bruno Sobral
Source of Support: NIH
Total Amount Awarded: $23,030,822
(VBI Portion: $12,783,340)
Total Award Period Covered: 10/01/09-09/30/14

Project Title: Probing oomycete-host interactions using an effector ORFeome and effector protein microarrays
PI: Brett Tyler
Co-PIs: J. Peccoud, J. McDowell
Source of Support: UC Davis (USDA)
Total Amount Awarded: $1,000,000
(VBI Portion: $222,974)
Total Award Period Covered: 05/01/10-04/30/13

Project Title: Rapidly adaptable nanotherapeutics
PI: Allan Dickerman
Source of Support: Techulon (DARPA)
Total Amount Awarded: $33,210
Total Award Period Covered: 06/01/12-03/31/13

Project Title: Rational design of high-affinity peptide drug candidates
PI: Enusha Karunasena
Source of Support: Lynntech, Inc. (NIH)
Total Amount Awarded: $415,192
Total Award Period Covered: 08/15/11-07/31/13

Project Title: Redox regulation of methanogenesis: A basic microbial process at the interface of climate change and renewable energy
PI: Biswarup Mukhopadhyay
Source of Support: NSF
Total Amount Awarded: $298,771
(VBI Portion: $196,693)
Total Award Period Covered: 06/01/10-05/31/13

Project Title: REU Site: Microbiology in the post-genome era
PI: Biswarup Mukhopadhyay
Source of Support: NSF
Total Amount Awarded: $288,285
Total Award Period Covered: 03/01/12-02/28/15

Project Title: REU Site: Microbiology in the post-genome era
PI: Biswarup Mukhopadhyay
Source of Support: NSF
Total Amount Awarded: $302,157
Total Award Period Covered: 06/01/09-05/31/13

Project Title: REU Site: Modeling and simulation in systems biology
PI: Reinhard Laubenbacher
Co-PIs: Christopher Lawrence
Source of Support: NSF
Total Amount Awarded: $386,157
Total Award Period Covered: 04/01/11-03/31/14

Project Title: REU Supplement to cross-layer optimization in cognitive radio networks in the physical interference model based on SINR constraints: Algorithmic foundations
PI: Anil Vullikanti
Source of Support: NSF
Total Amount Awarded: $16,000
Total Award Period Covered: 07/01/10-01/31/14

Project Title: Rigorous approaches for validation and verification of networked systems
PI: Madhav Marathe
Co-PIs: Christopher Barrett, Stephen Eubank, Henning Mortveit
Source of Support: DTRA
Total Amount Awarded: $1,190,969
Total Award Period Covered: 06/01/11-05/31/14

Project Title: SAIC fellowship in synthetic biology
PI: Jean Peccoud
Source of Support: SAIC, Inc.
Total Amount Awarded: $107,500
Total Award Period Covered: 09/01/08-05/31/13

Project Title: SDCI NMI new: From desktops to clouds -a middleware for next-gen network science
PI: Madhav Marathe
Co-PIs: Keith Bisset, Richard Alo, Geoffrey Fox, Edward Fox
Source of Support: NSF
Total Amount Awarded: $1,350,000
(VBI Portion: $512,254)
Total Award Period Covered: 09/01/10-08/31/13

Project Title: Stochastic models of the cell cycle regulation in eukaryotes
PI: John Tyson
VBI PI: Jean Peccoud
Source of Support: NIH
Total Amount Awarded: $2,563,991
(VBI Portion: $963,900)
Total Award Period Covered: 06/06/06-08/31/14
Effort: 1.00 months cy

Project Title: Supplement to NetSE Large Collaborative Research: Contagion in large socio-communication networks for the USA Science and Engineering Festival 2012
PI: Madhav Marathe
Source of Support: NSF
Total Amount Awarded: $25,459
Total Award Period Covered: 01/01/12-12/31/12

Project Title: Synthetic information systems for better informing public health policymakers
PI: Stephen Eubank, Madhav Marathe
Source of Support: NIH
Total Amount Awarded: $112,740
(VBI Portion: $78,391)
Total Award Period Covered: 08/01/12-07/31/13

Project Title: Synthetic information systems for better informing public health policymakers - supplement 1
PI: Stephen Eubank, Madhav Marathe
Source of Support: NIH
Total Amount Awarded: $228,510
(VBI Portion: $125,481)
Total Award Period Covered: 09/01/11-08/31/13

Project Title: Synthetic information systems for better informing public health policymakers
PI: Stephen Eubank, Madhav Marathe
Source of Support: NIH
Total Amount Awarded: $3,933,969
(VBI Portion: $3,315,332)
Total Award Period Covered: 05/01/11-04/30/16

Project Title: The role of the major *Alternaria* allergen Alt a 1 in airway inflammation
PI: Christopher Lawrence
Co-PIs: Brett Tyler
Source of Support: NIH
Total Amount Awarded: $428,536
Total Award Period Covered: 08/05/11-07/31/13

Project Title: The systems biology of iron homeostasis and the immune response to *Aspergillus*
PI: Reinhard Laubenbacher, Christopher Lawrence
Source of Support: NIH
Total Amount Awarded: $424,483
Total Award Period Covered: 07/01/12-06/30/14

Project Title: Virus tracker in a box
PI: Kristy Collins
Source of Support: RTI International
Total Amount Awarded: $19,998
Total Award Period Covered: 09/01/12-08/31/13

Project Title: Zinc-finger nuclease-mediated repair of disease-causing triplet CAG repeats
PI: David Mittelman
Source of Support: NIH
Total Amount Awarded: $424,483
Total Award Period Covered: 07/01/12-06/30/14


EDUCATION + OUTREACH

SUMMARY

The focus of the Education and Outreach Group at VBI encompasses programs that target the whole range of academic experience from K-12 to early career. We seek to strengthen the educational experience of the undergraduates, graduates, and early career scientists who work at VBI, building the infrastructure and networks to better track and recruit them and to develop more research opportunities. The aim of the K-12 program is to motivate children to become scientists by experiencing it firsthand as a world of discovery. We do this through programs like Kids' Tech University, which is spreading to other universities; the Genetics, Bioinformatics, and Computational Biology (GBCB) graduate program; and many other undergraduate and graduate opportunities.

PROGRESS REPORT

Kids' Tech University

Kids' Tech University is based on a program designed for children between the ages of 8-12 that started at the Eberhard Karls University, Tubingen, called *Die Kinder-Uni*. The founders wanted to provide children with a university experience characterized by interactive sessions held in university lecture halls and presented by university professors. The program proved so successful that *Die Kinder-Uni* are currently established at more than 275 universities in more than 8 European countries. In 2008, Dr. Reinhard Laubenbacher, the Director of the Education and Outreach Program at VBI, and Dr. Kristy Collins, a Senior Research Associate, seeded the program in the United States and it is thriving.

Today, Kids' Tech University can be found at Virginia State University (serves 100 children), Bowling Green State University (150), the Virginia Tech Southwest Center (100) and VT (450). Each year seems to bring another university on board. These programs relay heavily on foundation donations and registration fees. The Bowling Green State University program is currently being funded through a USDA grant.

VT's version is held on four Saturdays each spring. Last year's faculty presenters were Dr. Suzanne Weekes, associate head of the mathematical sciences department at Worcester Polytechnic Institute, Dr. Wu Feng, director of the Synergy Lab at VT, Dr. Stephen Eubank, deputy director of the Network Dynamics and Simulation Science Lab at VBI, and Dr. Brett Tyler, director of the Center for Genome Research at Oregon State University. In the afternoon of each event, VT professors, clubs, and local companies host hands-on exhibits for the children and their parents to explore. Please see http://kidstechuniversity.vbi.vt.edu/index.php for a complete description.

Genetics, Bioinformatics, and Computational Biology Graduate Program

This is an opportunity for graduate students to enter our labs in one of the first interdisciplinary PhD programs on the VT campus. Approximately 10 new students are accepted into GBCB each year and VBI faculty host the majority of them (40-45).

Because the VT campus is big, and funding is not always consistent, E&O spearheaded a drive to locate the GBCB website on the VBI servers to give greater control over the content to its stakeholders. With the help of our IT personnel, led by Jim Stoll, GBCB faculty can update their profiles with a simple logon; the GBCB chair and program manager have editing capabilities; and we've seeded the website with a basic but more organized feel. (See http://gbcb.vbi.vt.org). We feel websites are the window into the soul of a program and that this new look will help with recruitment.

Dr. Kathy O'Hara analyzed the GBCB program, identifying its strengths and weaknesses, and submitted a white paper to the VBI faculty for their assessment.

Undergraduate Summer Research programs

Several funding agencies support ten-week research experiences during the summer for undergraduates and VBI was privileged to host three by senior researchers Drs. Laubenbacher, Murali, and Mukhopadhyay. We also partnered with the Mathematical BioSciences Institute and hosted two of their students within Laubenbacher's group. Betsy Williams, VBI's Education and Outreach program coordinator, made the students feel at home by facilitating programmatic elements for the 24 students during their stay. Last year also saw a significant increase in the number of applicants and

consequently the quality of the students who were accepted. Kris Monger, E&O's webmaster, increased our recruitment efforts by joining various listservs and websites that advertise and encourage summer interns for undergraduates and it really paid off. This year Kris is monitoring our new online application system and reaching out to more minority serving institutions via the internet.

Undergraduate Scholars Program

Each semester 20-25 undergraduates train in our research labs. They come for course credit or simply to explore another part of their educational opportunities at VT. Research shows that students are more likely to stay in STEM fields if they have such research experiences, and we are proud to help make this happen. With the help of Kris Monger and Betsy Williams, we've streamlined the application process, created a database to track the students' progress, and will start some programmatic events in the Spring. Last year, Dr. Collins managed a day-long workshop, VBI Research Day, where VBI students from all levels could talk about their work with posters or Powerpoint.

High School Summer Internship

Last summer 16 high school students came on campus to discover how science works based on the theme "Our Green Future." This week-long internship, funded by the NSF, exposes high school students to current research in biology, the environment, climate, infectious disease, and bioinformatics. It consists of several activity labs, research-related tours, and a discussion/lecture series. Next summer (2013), the NIH will fund 12 high school students in a similar program which will focus on the "Modeling of Infectious Disease."

Building Networks: Recruiting, Research, Web Presence

Education and Outreach is never done alone. The job is defined by the creation of networks to further the educational mission of VBI, both on campus and off. To that end, Dr. Collins and Dr. O'Hara attended professional conferences and workshops to establish new and strengthen old paths for the recruitment of students. Here is a list of past activities:

Collins K. Teacher Workshop- "Patterns are Everywhere! How and Why?", Bowling Green State University Kids' Tech University, Bowling Green, OH. January 21, 2012.

Collins K. ANDRILL's (Antarctica Geological Drilling) Environmental Literacy Framework, Summer STEM Institute Abingdon, VA. July 18, 2012.

Collins K. Antarctica's Climate Secrets, Buchanan County Middle School Teachers, Grundy, VA. November 12, 2012.

Collins K. Extracting DNA from Strawberries, Virginia 4-H Volunteers, Blacksburg, VA. November 17, 2012.

Collins K. Kids' Tech University, European Union of Children's University, Oporto, Portugal. November 28-December 1,2012.

O'Hara, K. Panel member and recruiter, Undergraduate Research Capstone Conference, Mathematical Biosciences Institute, Columbus, OH. August 13-17, 2012.

O'Hara, K. Recruiter and Math Poster Judge, Annual Biomedical Research Conference for Minority Students, San Jose, CA. November 7-10, 2012.

O'Hara, K. Participant and recruiter, National McNair Scholars research conference and Graduate School.