COLUMBIA UNIVERSITY
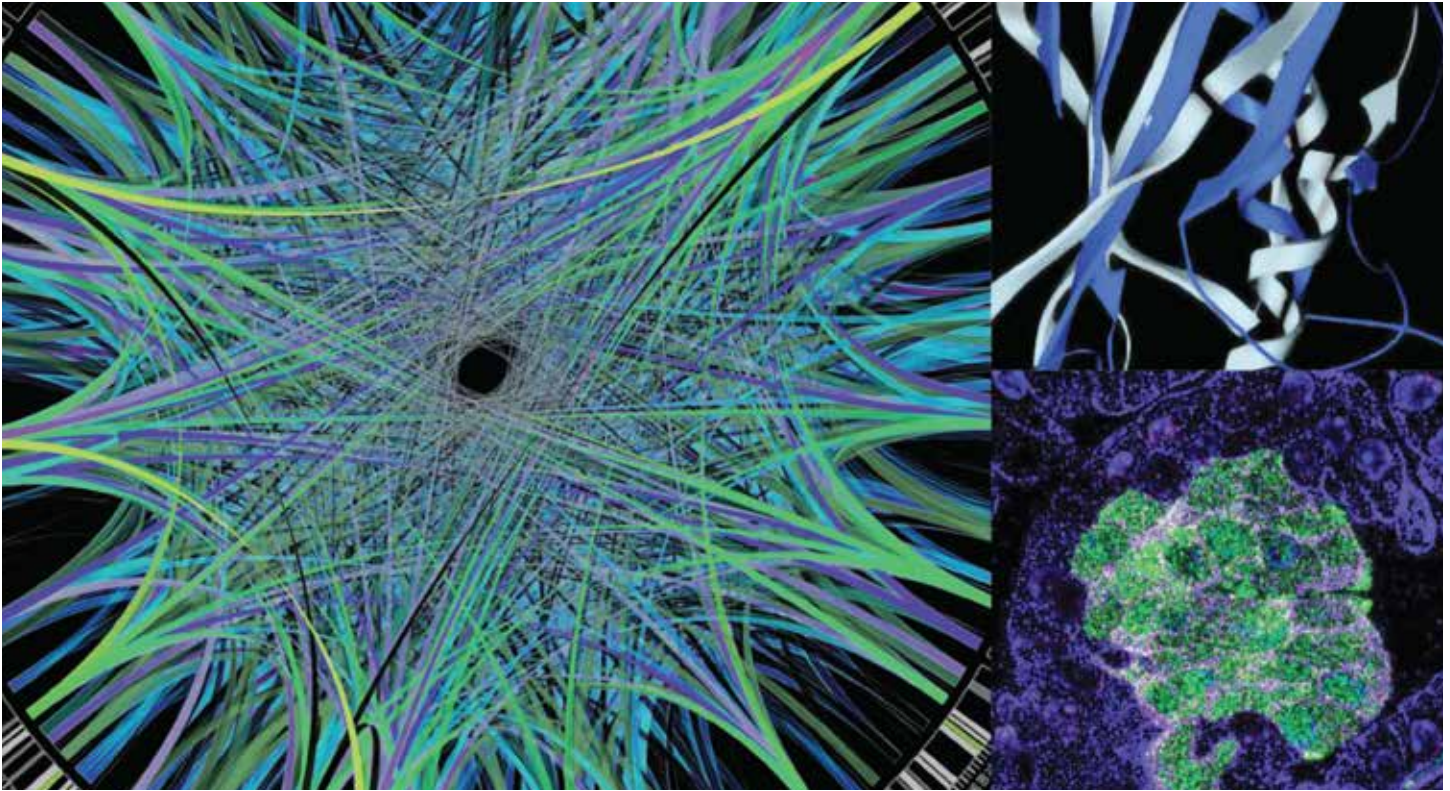## Department of Systems Biology

A Vision for the Future of Biomedical Research

In the history of science, quantitative models have triggered golden eras of discovery in many scientific disciplines.

Consider, for example, the ways in which modeling and computational simulation have improved our understanding of high-energy particles in physics, weather patterns in climatology, and the flow of capital in economics. In these fields and others, methodologies derived from the quantitative sciences have leveraged large-scale empirical data sets and provided powerful tools for generating and testing hypotheses. In each case, the subsequent availability of high-performance computing platforms led to further leaps forward, allowing simulations of models that were too complex to be solved analytically.

In biology and medical research, recent developments suggest that we have reached a similar turning point. Over the past decade, new technologies have made it possible to measure, on a genome-wide scale, many variables that determine cellular activity. This has led to an explosion in the amount of data that are now available — far too much to

analyze using traditional methods. In response, scientists trained in systems and computational biology have begun to play an increasingly important role. Using algorithms and complex models of cellular regulation running on powerful supercomputers, these scientists can now mine the data for insights into both basic biology and the biology of human health. The computational models that they develop have generated new hypotheses and have led to many surprising discoveries. Perhaps most importantly, these approaches have improved the efficiency of biological research, pointing the way toward experiments that are most likely to succeed.

When the human genome sequence was published in 2001, many hoped that we had uncovered the Rosetta Stone that would enable us to conquer disease. A number of revolutionary advances since then have made it clear, however, that the genome is just a list of parts. How these parts can be combined or assembled to form distinct living systems is still largely unknown. Scientists rarely talk anymore about finding "the gene that causes x," but more often about how physiologic traits and diseases emerge from the interaction of multiple gene factors and their cellular environment. For these reasons, the coming decades of biolo-

At Columbia, systems biology integrates various types of information in order to understand how cells behave at the level of genes, proteins, and cells. Research can include developing models of interactions between proteins and genes (far left), computationally predicting protein structure (top right), and performing experimental tests in the laboratory to validate these and other types of computational models. (bottom right).

Andrea Califano
Chair,
Department of
Systems Biology
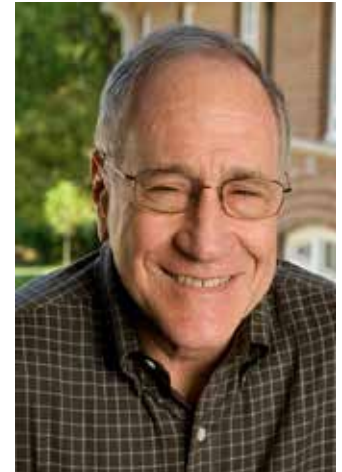
Barry Honig
Director,
Center for
Computational Biology
& Bioinformatics

gy will require more than reading the code written in the genome; this information will need to be integrated with data about protein structure, gene expression levels, and chemical modifications that affect the activities of both genes and proteins. Furthermore, these rich sources of data will need to be interpreted through the lens of what we call the "regulatory logic" of the cell; that is, the multilayered networks of molecular interactions that are specific to particular cell types and contexts. This is an ambitious agenda, but already biologists have begun to map these networks in an effort to understand how all the pieces fit together.

Systems biology is the discipline that is best poised to understand biology from the perspective of complexity. It appreciates and embraces the challenge of determining how large numbers of gene products work together to produce observable traits and diseases, rather than studying one gene at a time. It has also produced a growing repertoire of experimental and computational tools designed to dissect complexity at different levels of biological activity. Moreover, in just a little over a decade, systems biology has begun generating important discoveries that are relevant for both basic biology and biomedical research.

At Columbia, discoveries using these perspectives have even opened the way for multiple clinical trials, which we anticipate could offer new opportunities for more effective therapeutics in the coming age of precision medicine.

With its decision to create a Department of Systems Biology in 2013, Columbia University took a bold step to embrace this new approach. This commitment reflects the strength of the systems biology research community at Columbia as well as the ways in which the perspective and tools that systems biology offers are informing research done across the university. In its strategic plan, *2020 Vision*, the College of Physicians and Surgeons identified systems biology as one of the "scientific priorities that will define the future of health and biomedical science." We are very excited about the outstanding community of investigators who are pushing the discipline forward at Columbia, and are proud of the many productive collaborations that researchers in our community have undertaken, as well as their resulting achievements.

Only time will tell what role systems biology will play in the next generation of biological research. But wherever it goes, we are excited that Columbia will be there to help lead the way.

3

Our Collaborative Research Community

The mission of the Columbia University Department of Systems Biology is to develop new methods for understanding the biological world from a systems perspective, particularly at the genomic and molecular levels. Our faculty shares a common interest in combining high-throughput experimentation, quantitative analysis, and innovative technology development. Researchers use computational approaches and laboratory experimentation in an iterative way, developing predictive models of biological systems, and then validating them in the laboratory.

With more than 30 faculty members, a community of more than 300 scientists working in department labs, and an outstanding record of publications, the Columbia University Department of Systems Biology is among the largest and most accomplished programs in this field.

The Department is highly multidisciplinary, and collaborates with investigators in other institutes at Columbia University and at other universities. Methods based in systems and computational biology are providing insights into human diseases that would be impossible using other methods.

Researchers in the Department of Systems Biology develop innovative cutting-edge sequencing and microscopy methodologies as well as software tools and databases for the study of various facets of biology. Through an open source, web-based platform called geWorkbench, we make these resources available to researchers anywhere.

Columbia University investigators have also played a leadership role in promoting the exchange of ideas within the global systems biology community by organizing key conferences and meetings.

Assistant Professor Peter Sims (above) is developing new technologies for genome-wide sequencing, one individual cell at a time. Staff In the Columbia Genome Center (opposite page, top) use robotics and highly precise automation technologies to run experiments on up to millions of samples in parallel. Researchers in the laboratory of Associate Professor Dana Pe'er (opposite page, bottom) have been developing computational approaches for analyzing large, complex collections of biological data, particularly for purposes of cancer research. Photos: Lynn Saville and Amelia Panico.

Faculty

Cory Abate-Shen
Cancer biology, mouse models

Dimitris Anastassiou
Computational biology, cancer biomarkers

Harmen Bussemaker
Computational biology, gene regulation

Andrea Califano
Cancer systems biology, precision medicine

Virginia Cornish
Chemical biology, synthetic biology

Adolfo Ferrando
Cancer genomics

Aris Floratos
Bioinformatics, software development

Oliver Hobert
Gene regulation, cellular differentiation

Barry Honig
Computational structural biology

Tuuli Lappalainen
Computational and population genomics

Kam Leong
Nanotechnology, biomedical engineering

Richard Mann
Regulatory genomics, development

Dana Pe'er
Computational biology, single-cell analysis

Itsik Pe'er
Computational biology, population genetics

Molly Przeworski
Population genetics

Raul Rabadan
Computational genomics, applied mathematics

Rodney Rothstein
Genetics, genome stability

Sagi Shapira
Immunology, host-pathogen dynamics

Lawrence Shapiro
Structural biology, protein binding

Michael Shen
Stem cell and cancer biology

Yufeng Shen
Bioinformatics, genome sequence analysis

Peter Sims
Single-cell sequencing technologies

Brent Stockwell
Chemical biology and screening

Milan Stojanovic
Biomedical engineering

Gustavo Stolovitzky
Computational biology, DREAM Project

Nicholas Tatonetti
Biomedical informatics

Saeed Tavazoie
Gene regulation, adaptive evolution

Dennis Vitkup
Metabolic networks, probabilistic modeling

Harris Wang
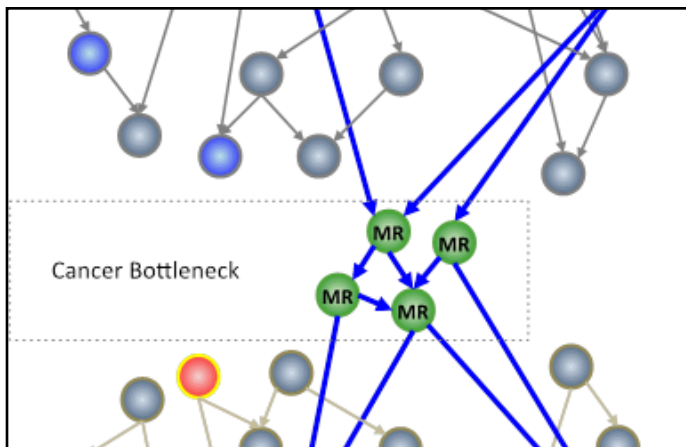Synthetic biology, microbiome engineering

Chris Wiggins
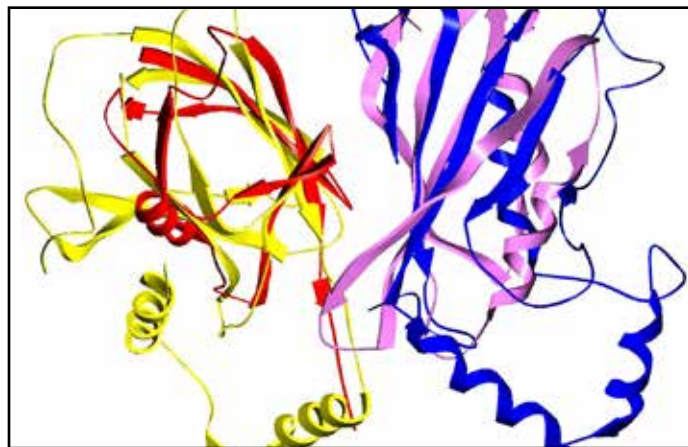Machine learning

Chaolin Zhang
RNA networks, neurodegenerative disease

Research Briefs

N-of-1 Clinical Trials for Cancer Started

Uniting Structural and Systems Biology

A team led by Andrea Califano, in collaboration with physicians at Columbia University Medical Center has launched an innovative new approach to clinical trials aimed at improving precision medicine for cancer. Instead of focusing on mutated genes associated with cancer, these "N-of-1 trials" use systems biology methods developed in the Califano Lab to analyze the molecular regulatory networks that drive an individual patient's tumor cells. By reverse engineering these regulatory networks, Califano identifies genes and gene pairs — called "master regulators" — that are critical bottlenecks within these networks and are essential for the survival of the tumor. If FDA-approved drugs that inhibit these genes exist, they are then tested in a mouse model of the patient's tumor and, if successful, may be investigated further in a more traditional clinical trial. Research has indicated that this systems biology-based approach could offer a powerful new paradigm for cancer diagnosis and for identifying effective precision treatment strategies. The scientists also expect that master regulators identified in one tumor could play important roles in others. As the results of N-of-1 studies accumulate, investigators will begin to assemble a more precise understanding of cancer subtypes. In this way, N-of-1 clinical trials will focus scientsits' attention on the best therapeutic opportunities, accelerating future cancer research.

When the Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet) was founded in 2005, one of its goals was to integrate the methods of structural biology with those of systems biology. Determining whether two proteins were structurally capable of interacting, researchers hoped, would both improve the predictive value of models of molecular interactions by giving another layer of complementary evidence, and lead to new types of predictions that could not be made using other methods. In a paper published in *Nature*, Barry Honig, Andrea Califano, and colleagues reported on a novel approach for using information about protein structures to make predictions about protein-protein interactions on a genome-wide scale. Their approach capitalizes on methods the Honig lab has developed over the last 15 years for inferring a protein's activity by predicting structural similarities between it and other proteins with known local structures and functions. This work culminated in the development of an algorithm called Predicting Protein-Protein Interactions (PrePPI), which has helped to predict 300,000 interactions in humans that are highly likely to occur in nature, a 10-fold increase over what had previously been identified. "This approach is new — both for structural and systems biology," Honig says. "We can now expand the utility of structural information to generate predictive models of protein-protein interactions."[1]

6
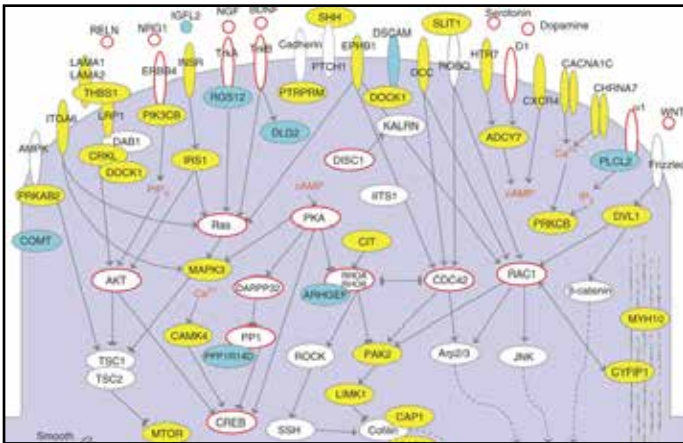
Connections Found between Genetic Networks for Schizophrenia and Autism

Dennis Vitkup and colleagues performed network-based analyses of rare genetic mutations to map the gene networks that lead to schizophrenia. When they compared one schizophrenia network to a previously discovered autism network, they discovered that different copy number variants in the same genes can lead to either schizophrenia or autism. The overlapping genes are important for axon guidance, synapse function, and cell migration — processes in the brain that have been shown to play a role in the development of these two diseases. These gene networks are particularly active during prenatal development, suggesting that the foundations for schizophrenia and autism are laid very early in life. Dr. Vitkup believes that there may be many more genes to be found that are tied to schizophrenia, but predicts that they will function within the networks that his team has described. "Until a few years ago," he explained, "people were looking for just a handful of genes responsible for autism and schizophrenia, so the idea that many hundreds of genes are involved is a big change in thinking... Our study and the studies of our collaborators suggest that in the search for the causes of complex genetic disorders, it will be more productive to look for common pathways and gene circuits than for a handful of causal genes. This type of network analysis gives us a way to begin to make sense of what's happening."[2]
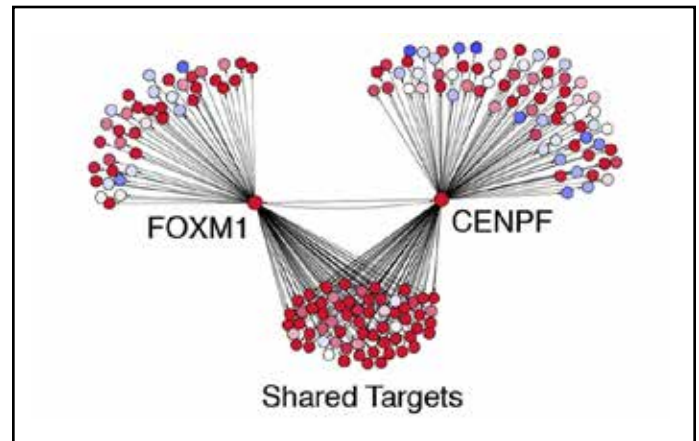
Synergy between Two Genes Drives Aggressive Prostate Cancer

Two genes — FOXM1 and CENPF — have been found to work together to cause the most aggressive form of prostate cancer. In a study led by Cory Abate-Shen, Michael Shen, and Andrea Califano and published in *Cancer Cell*, investigators devised a novel experimental approach that used computational methodologies to compare the gene regulatory networks that drive prostate cancer in humans with those in a genetically engineered mouse model of the disease. Using the high-performance computing cluster housed in the Department of Systems Biology, the analysis determined that aberrant expression of either FOXM1 or CENPF does not activate programs associated with the worst hallmarks of the disease. When acting together, however, the two genes can wreak havoc in the cancer cell and turn it into a very aggressive tumor. After this discovery the researchers analyzed prostate cancers from a group of more than 900 patients who had undergone prostate removal surgery, and found a striking correlation between the co-expression of FOXM1 and CENPF and the poorest disease outcome. The researchers also showed that simultaneously silencing the two genes inactivated pathways known to be hallmarks of aggressive prostate cancers. The finding suggests the possibility that combined therapeutic targeting of both FOXM1 and CENPF could arrest human disease, and is being investigated further in a clinical trial.[3]

continued from previous page

Distinguishing Patterns of Tumor Evolution in Chronic Lymphocytic Leukemia

As biologists have gained a better understanding of cancer, it has become clear that tumors are often driven by a series of genetic changes that correspond to particular stages of cancer progression. In this sense, a tumor is constantly evolving, and so one key question is how to disentangle the order in which mutations occur in order to understand how tumors change over time. A team of investigators led by Raul Rabadan developed a new strategy, called tumor evolutionary directed graphs (TEDG), which integrates next-generation sequencing data from tumor samples from a large number of patients. Using TEDG to analyze cancer cells in patients with chronic lymphocytic leukemia, the team discovered a consistent pattern in which mutations in CLL follow a branching pattern. Interestingly, the study also showed that each individual case of CLL can follow one of two distinct evolutionary trajectories, indicating that there may be at least two different molecular subtypes of the disease. The investigators believe that considering tumor evolution could provide a more effective way of categorizing tumors than cancer genetics approaches that rely on the presence of a single dominant clone.[4]

Bacterial Fitness in the Gut Microbiome

Little is known about why certain bacterial strains in the human gut outcompete others in the same environment. Harris Wang, in collaboration with researchers at Harvard University, developed a new method called TFUMseq to identify genes within these bacteria that give them an evolutionary fitness advantage. In an initial test, they took fragments from the genome of the bacteria *Bacterioides thetaiotaomicron (Bt)* and inserted them into *E. coli* bacteria. They then gave food containing these engineered bacteria into mice with no bacteria in their gut. As the colony grew, they extracted bacteria from the mice's feces at multiple time points, looking for the relative abundance of each of the *Bt* gene fragments, and used a new computational approach to identify which fragments conferred the greatest selection advantage. The approach could offer strategies for engineering probiotic bacteria that could more effectively deliver beneficial gene products.[5]

Method for Analyzing Single-Cell Data Identifies AML Stem Cells

There is enormous genetic heterogeneity among the millions of cells that make up an individual tumor. Even within an individual, these cells diversify into subpopulations with unique properties, though distinguishing these subpopulations is incredibly difficult. Of particular interest are cancer stem cells, which are typically resistant to cancer therapies and lead to relapse and recurrence of cancer following treatment. In a paper published in *Cell*, investigators in the laboratories of Dana Pe'er and Garry Nolan (Stanford U.) used an experimental technology called mass cytometry alongside a new computational algorithm, PhenoGraph, for analyzing the resulting data. The team classified millions of individual blast cells responsible for acute myelogenous leukemia (AML) into subpopulations with distinct phenotypes. Their method revealed a pattern of cell signaling that was indicative of a primitive cell state that could distinguish AML-driving cancer stem cells. Additional investigation also revealed a gene expression signature that was predictive of patient survival. The method is not only applicable to AML, but could offer a general strategy for identifying subpopulations of cells with distinctive phenotypes.[6]

References

[1] Zhang QC et al. **Structure-based prediction of protein-protein interactions on a genome-wide scale**. Nature. 2012 Oct 25;490[7421]:556-60.

[2] Gilman SR et al. **Diverse types of genetic variation converge on functional gene networks involved in schizophrenia**. Nat Neurosci. 2012 Dec;15[12]:1723-8.

[3] Aytes A et al. **Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy**. Cancer Cell. 2014 May 12;25[5]:638-51

[4] Wang J et al. **Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia**. Elife 2014 Dec 11;3.

[5] Yaung SJ et al. **Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics**. Mol Sys Bio 2015 Mar 11;11[3]:788.

[6] Levine JH et al. **Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis**. Cell. 2015 Jul 2;162[1]:184-97.

Centers of Excellence
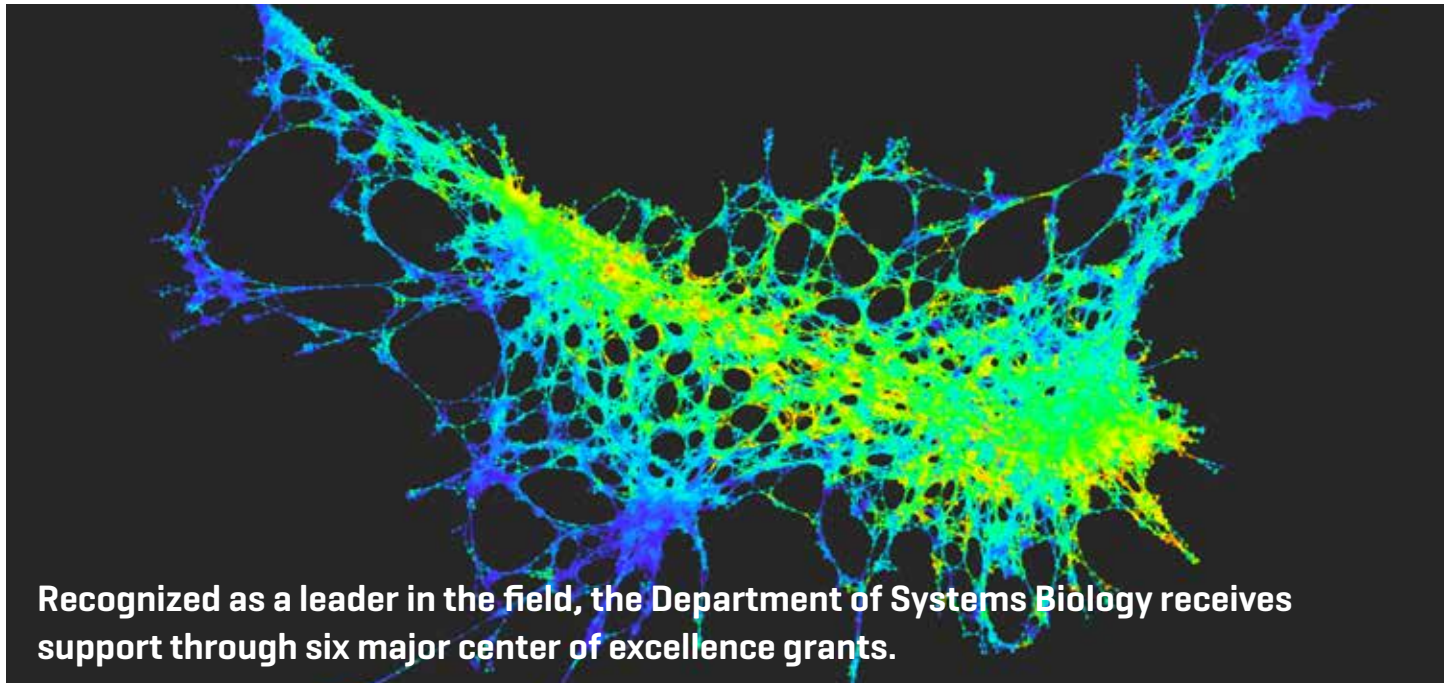
Recognized as a leader in the field, the Department of Systems Biology receives support through six major center of excellence grants.

Center for Multiscale Analysis of Genetic and Cellular Networks (MAGNet)

One of 8 National Centers for Biomedical Computing, a center of the NCI's Integrated Cancer Biology Program, and one of 12 interdisciplinary Centers for Cancer Systems Biology, MAGNet develops computational tools for identifying molecular interactions in the cell and how they produce cellular phenotypes.

Library of Integrated Network-Based Cellular Signatures (LINCS)

This initiative is developing a library of observations of molecular activities that occur when specific cell types cells are exposed to specific chemicals. We are profiling molecular signatures that occur when cells are exposed to multiple chemical compounds in synergy, and using systems approaches to explain mechanisms of drug activity.

International Serious Adverse Event Consortium (iSAEC)

iSAEC's mission is to identify DNA variants that confer a risk for drug-induced adverse events, knowledge that will help to predict risk to patients when taking specific medications and to design treatments accordingly.

Center for Topology of Cancer Evolution and Heterogeneity

A member of the National Cancer Institute's Physical Sciences in Oncology Program, this multidisciplinary center is developing methods that combine mathematics and new experimental techniques for the study of cancer. It also organizes the New York Metropolitan Area Discussion Group in Mathematics and Oncology.

Cancer Target Discovery and Development Center (CTD^2)

This program of the National Cancer Institute develops new ways to use the rapidly growing body of cancer genomics data to systematically and efficiently develop new precision treatments for cancer. We are using systems biology approaches to study tumor progression and resistance to chemotherapy.

The graph above was generated in the laboratory of Associate Professor Raul Rabadan. As a lead investigator in the Center for Topology of Cancer Evolution and Heterogeneity he is developing novel applications of advanced mathematics to predict how small populations of genetically distinct cells within solid tumors change over time and promote tumor growth.

Training a New Generation of Investigators

The Department of Systems Biology and its Center for Computational Biology and Bioinformatics provide graduate education that prepares students to become leaders in this exciting and rapidly growing field.

Our leadership role in this area has been recognized by the NIH in the form of continued funding of its training grant in computational biology. Education at Columbia stresses the importance of systems-wide analyses and modeling, high-throughput experimentation, computational data integration, and innovative technology development. Students learn to develop and apply powerful systems and computational biology approaches for investigating challenging biological and biomedical problems.

Our students arrive with diverse interests in the biological sciences, computer science, chemistry, mathematics, physics, engineering, and other related fields. During graduate studies, students engage in research that often results in high-profile publications.

Because of the increasing importance of integrated, data-intensive research to the future of the biological sciences, graduates with expertise in systems and computational biology are currently in high demand. Upon graduation, our alumni have a strong track record of launching successful careers in both academia and industry.

Professor Virginia Cornish (right) combines methods in organic chemistry and DNA technology to expand the synthetic capabilities of living cells. Photo: Lynn Saville.

Integrated Facilities for Biological Research

Studying biology on a systems-wide scale requires the latest experimental and computational technologies. The Department of Systems Biology has developed a robust infrastructure that combines the two.

In cooperation with the Herbert Irving Comprehensive Cancer Center (HICCC), the Department of Systems Biology manages the **Judith P. Sulzberger Columbia Genome Center**, a state-of-the-art environment for conducting high-throughput biomedical research. Our **Genome Sequencing and Analysis Facility** is a certified Illumina CSPro Center, producing high-quality and cost-effective next-generation genome, exome, and RNA sequencing data. Our clients and collaborators include researchers in the Columbia University community, as well as investigators at other universities and in the biotechnology and pharmaceutical industries. Our **High-Throughput Molecular Screening Facility** operates as a collaborative resource center, working with investigators throughout the Columbia Research Community. We design, optimize, and perform assays that use robotic automation, multi-label detection, and chemical screening libraries to run millions of experiments in parallel. The Genome Center also maintains instrumentation for high-throughput high-content microscopy, as well as multiplexed measurement of gene and protein expression.

Complementing these experimental facilities, the **Department of Systems Biology Information Technology Facility (DSBIT)** maintains the large-scale, high-performance computing platforms required to analyze, visualize, and interpret massive amounts of data. For nearly a decade, DSBIT has been dedicated to providing computing services for academic research, and currently supports more than 80 diverse teams of investigators. Housed in the Irving Cancer Research Center building at Columbia University's medical campus, DSBIT manages a 3,000 sq. ft. data center that is customized for analyzing genomic and bioinformatics data. DSBIT's high performance compute cluster (HPC) bristles with 6,384 CPU cores, 23TB of RAM, and 148 NVIDIA GPUs, providing an additional 75,776 CUDA cores. A 1.6 PB enterprise-grade storage system is tightly integrated with the HPC cluster to support big data analyses. All of the HPC cluster is interconnected with a 10GB mesh network. For highly coupled parallel computations a portion of the cluster is also equipped with a 40GB QDR infiniband network. DSBIT also offers complementary web and database hosting, server virtualization, data center colocation, and desktop support.

The Columbia Genome Center's Genome Sequencing and Analysis Facility [left] performs high-quality next-generation genome, exome, and RNA sequencing. The Department of Systems Biology's computing center [right] provides powerful computational resources for data intensive research. Photos: Lynn Saville.

Remembering Judith Sulzberger

Photo: Carlos René Perez

Judith P. Sulzberger (1923-2011) was a visionary physician-philanthropist who applied her medical knowledge to advance research at the Columbia University College of Physicians & Surgeons (P&S) and elsewhere. A former member of the Board of Directors of the *New York Times*, she served on the CUMC Board of Visitors, the P&S Alumni Steering Committee, and the Health Sciences Advisory Council.

As genome studies began to grow in importance to biomedical research, Dr. Sulzberger provided the guidance, the seed money, and ongoing support for what would become the Judith P. Sulzberger Columbia Genome Center. Her philanthropy also supported the launch of the Columbia Initiative in Systems Biology (now the Department of Systems Biology), as well as other initiatives at Columbia.

About Columbia University Medical Center

Photo: Eileen Barroso

Columbia University Medical Center provides international leadership in basic, preclinical, and clinical research; medical and health sciences education; and patient care. The medical center trains future leaders and includes the dedicated work of many physicians, scientists, public health professionals, dentists, and nurses at the College of Physicians and Surgeons, the Mailman School of Public Health, the College of Dental Medicine, the School of Nursing, the biomedical departments of the Graduate School of Arts and Sciences, and allied research centers and institutions.
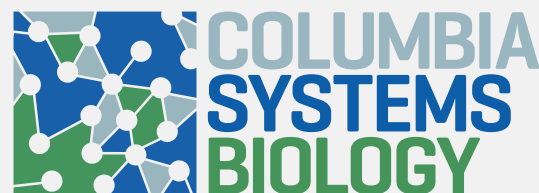
Columbia University Medical Center is home to the largest medical research enterprise in New York City and State and one of the largest faculty medical practices in the Northeast.

Contact Us

Columbia University Department of Systems Biology
Irving Cancer Research Center
1130 St. Nicholas Avenue, 8th Floor
New York, NY 10032
Phone: 212-851-5208

To learn more about our research and programs, visit us online at systemsbiology.columbia.edu.

Front cover image: Global mapping of cancer gene expression changes to the human metabolic network. Image by Jie Hu and Dennis Vitkup using the iPath tool.