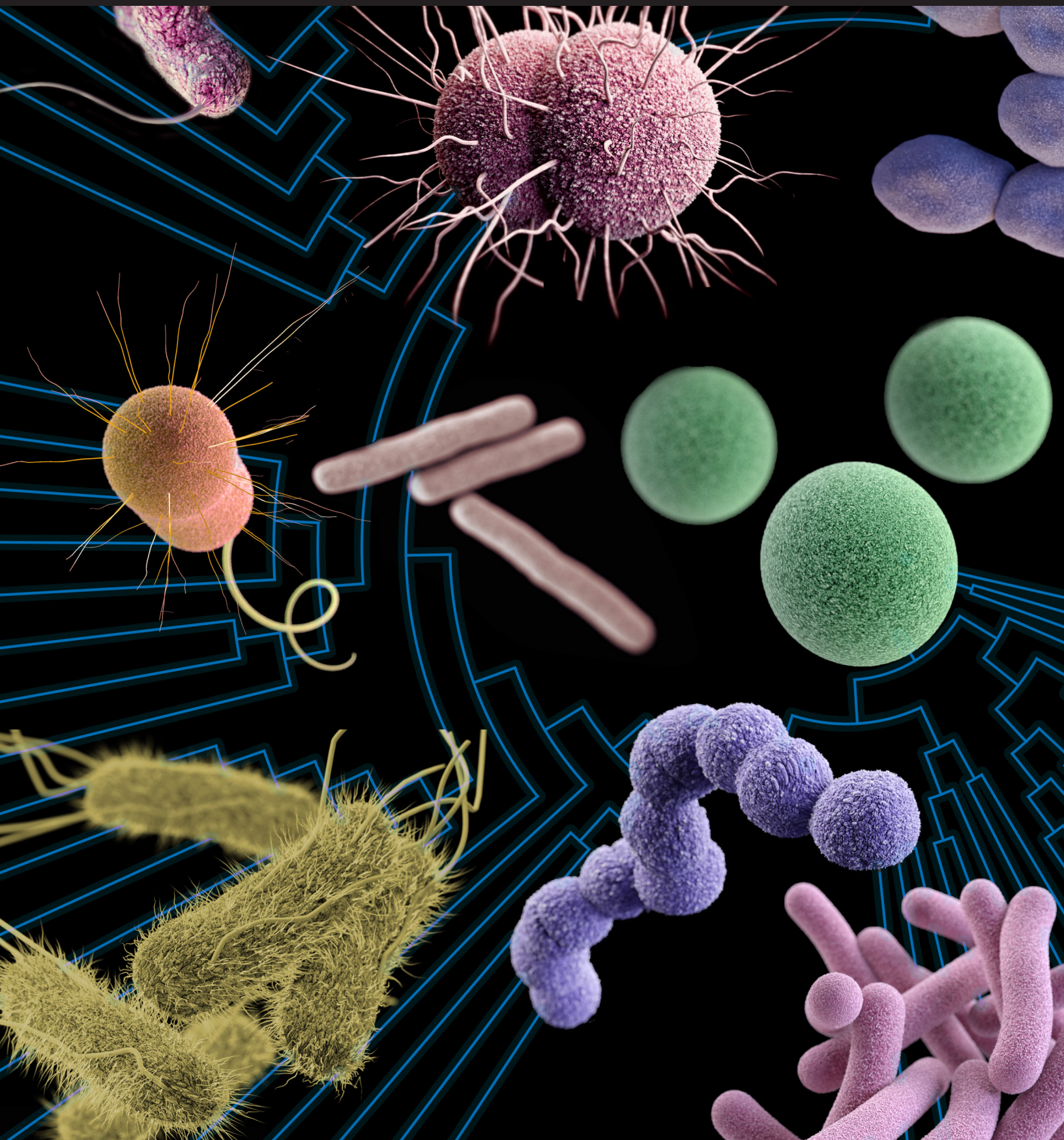
COLUMBIA UNIVERSITY
Department of Systems Biology

Research Highlights
June 2015



Functional Metagenomics Enables First Study of Bacterial Fitness in the Gut Microbiome

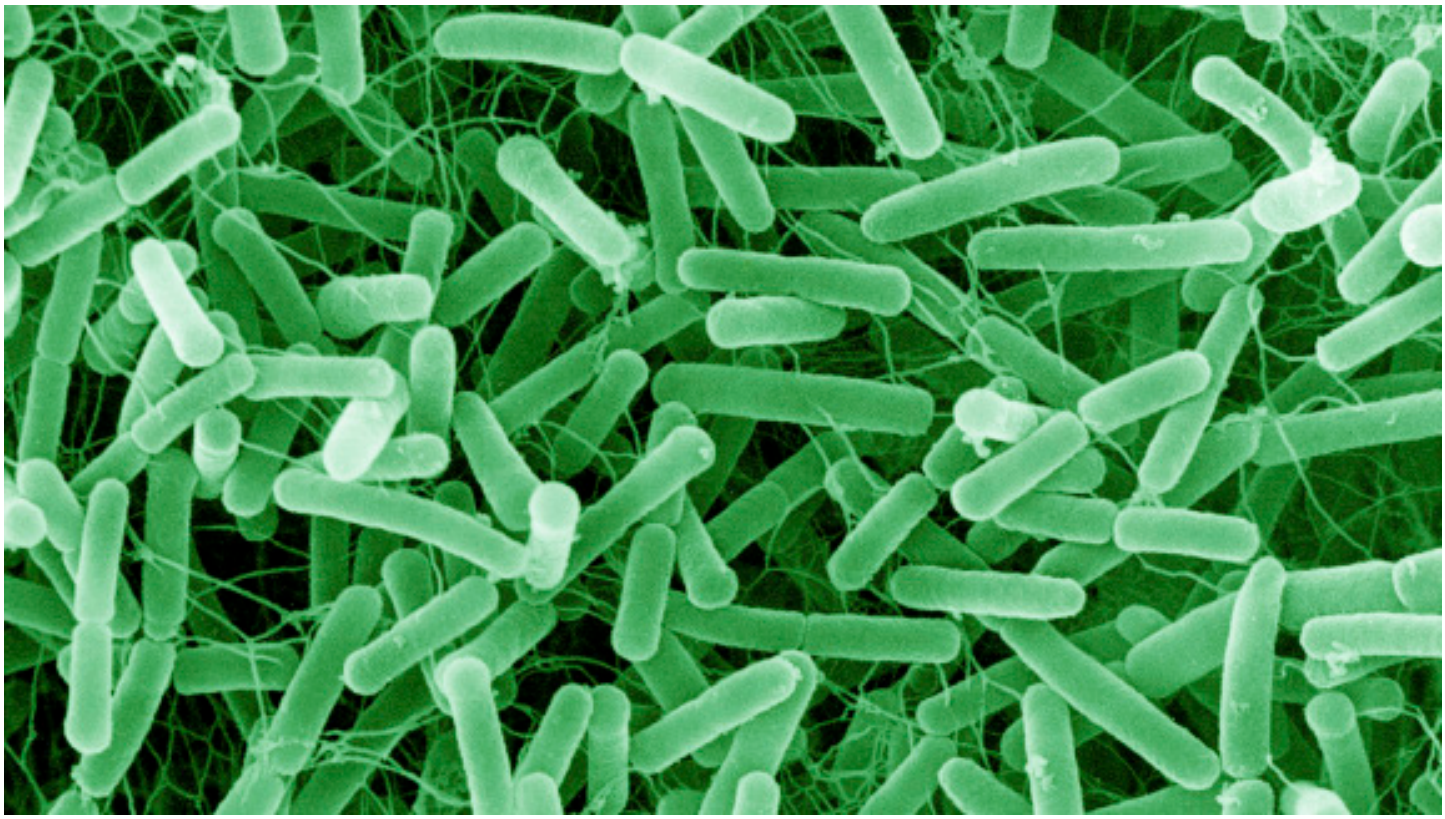


Photo by David Gregory and Debbie Marshall, Wellcome Images.

Recent deep sequencing studies are providing an increasingly detailed picture of the genetic composition of the human microbiome, the diverse collection of bacterial species that inhabit the gut. At the same time, however, little is known about the dynamics of these colonies, particularly why certain microbial strains outcompete others in the same environment. In a paper published in the journal *Molecular Systems Biology*, Department of Systems Biology Assistant Professor Harris Wang, in collaboration with Georg Gerber and researchers at Harvard University, report on their development of the first method for using functional metagenomics to identify genes within commensal bacterial genomes that give them an evolutionary fitness advantage.

In the initial test of their approach, called Temporal Functional Metagenomics Sequencing (TFUMseq), 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 *E. coli* to mice bred to have no bacteria in their gut. This provided a kind of sterile, blank slate for the new bacterial colony to grow. The team then extracted bacteria from the mice's feces at multiple time points across a span of 28 days. For each sample, they sequenced the bacteria and looked for the relative abundance of bacteria containing each of the Bt gene fragments. They then used a new computational analytic approach based on information theory to analyze these measurements and provide an index of which fragments conferred the greatest selection advantages. Because measurements were taken at multiple time points, they also showed how the competitive environment within the gut changed over time, enabling them to make accurate hypotheses of the underlying molecular activity.

Bacterioides thetaiotaomicron is a very common component of the microbiome that grows reasonably well in a laboratory setting. However, many other bacteria that grow in the gut do not, making them extremely difficult to study. Combining function-

Related publication:
Yaung SJ, Deng L, Li N, Braff JL, Church GM, Bry L, Wang HH, Gerber GK. **Improving microbial fitness in the mammalian gut by in vivo temporal functional metagenomics**. Mol Sys Bio 2015 Mar 11;11(3):788.

al genomics, deep sequencing, and computational methods in this way, the authors posit, offers a powerful new tool for doing so. In the experiments described in the paper they discovered several previously unknown details about the genes that enable Bt to multiply in the gut, including nutrient sources that are critical for Bt metabolism and enable the microbe to thrive.

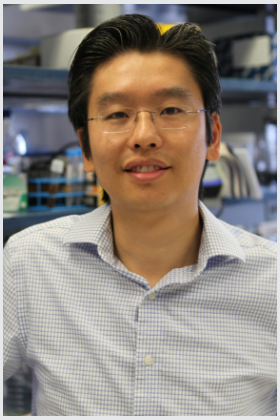
The knowledge that their approach generates, the researchers anticipate, could offer strategies for engineering probiotic bacteria that retain a fitness advantage after being ingested, enabling them to persist in the gut longer and to be used more effectively to deliver other beneficial gene products. The authors also suggest that their approach could help to engineer bacteria to have a fitness advantage in the same niche as known pathogens. In this way, the engineered bacteria would outcompete the pathogens for available sources of energy and reduce their deleterious effects.

In a "News & Views" review that appears in *Molecular Systems Biology* along with the paper, Jeremiah Faith discusses the potential implications of the method:

> The framework presented by Yaung et al for understanding the functional capacity of gut microbes drives to the heart of the most fundamental requirement of any organism in the microbiome, namely its ability to stably colonize a host in the context of the rich genetic diversity of competing organisms and host factors… [I]dentifying the genetic elements that confer fitness advantages during colonization is essential for understanding the final assembly of each individual's microbiota. Advantages of TFUMseq compared to existing technologies include its simplicity and, more importantly, its potential to be applied to forward engineer bacterial strains with enhanced colonization abilities in defined contexts, such as specific dietary interventions.

In the future, Wang says, the researchers plan to test their method in mice that already have an existing microbiota, instead of being germ-free. By studying how different microbiota change the competitive profile of different bacterial genes, they anticipate being able to learn more about factors that improve competition, as well as about the composition of the different gut environments themselves.

Harris Wang Named Office of Naval Research Young Investigator



Harris Wang has been selected for the Office of Naval Research 2015 Young Investigators Program. With the support of this highly selective award, Dr. Wang will extend his research in the field of synthetic biology to develop new technologies for engineering the gut microbiome. These new methods, Wang anticipates, could provide new ways of designing communities of different microbial species and ultimately modulating interactions between the gut, the immune system, and the brain.

Wang's project builds on recent observations suggesting that interactions between the brain and bacteria in the gut affect human physiology through the enteric nervous system (ENS). Changes in the composition of the microbiome, for example, can lead to various acute and chronic diseases that can affect wellbeing. In addition, emerging evidence suggests that neuropsychiatric factors such as stress can affect the composition and function of bacteria in the gut. This bidirectional communication can regulate how we physically experience emotions and even make decisions.

Within the context of synthetic biology, Wang's research addresses three significant limitations in scientists' current ability to engineer gut microbial communities: the challenge of making site-specific targeted changes to bacterial genomes, the limited repertoire of plasmids and mobile vectors for delivering modified DNA into target bacteria, and the lack of methods to introduce DNA into multiple species of bacteria at the same time. These problems need to be solved if synthetic biology is to become effective in engineering bacterial communities. The Wang Lab aims to address these challenges by developing a new, targeted genome engineering approach, expanding the library of useful plasmid vectors, and exploring new DNA transformation methods in microbial consortia.

Under the auspices of the Office of Naval Research grant, the ability to engineer the microbiome could lead to new strategies for improving resilience to stress and operational performance in naval personnel. If successful, Wang's project could lead to much wider applications, however, providing a transformative and foundational platform for engineering communities of microbes.

The Exposome: Relating Environmental Factors to Human Disease

Although genomics has dramatically improved our understanding of the molecular origins of certain human genetic diseases, our health is also influenced by exposures to our surrounding environment. Molecules found in food, air and water pollution, and prescription drugs, for example, interact with genetic, molecular, and physiologic features within our bodies in highly personalized ways. The nature of these relationships is important in determining who is immune to such exposures and who becomes sick because of them.

In the past, methods for studying this interface have been limited because of the complexity of the problem. After all, how could we possibly cross-reference a lifetime's worth of exposures with individual genetic profiles in any kind of meaningful way? Recently, however, an explosion in the generation of quantitative data related to the environment, health, and genetics — along with new computational methods based in machine learning and bioinformatics — have made this landscape ripe for exploration.

At this year's South by Southwest Interactive Festival in Austin, Texas, Department of Systems Biology Assistant Professor Nicholas Tatonetti and his collaborator Chirag Patel (Harvard Medical School) discussed the remarkable new opportunities that "big data" approaches offer for investigating this landscape.

Driving Tatonetti and Patel's approach is a concept called the exposome. First proposed by Christopher Wild (University of Leeds) in 2005, an exposome represents all of the environmental exposures a person has experienced during his or her life that could play a role in the onset of chronic diseases. Tatonetti and Chirag's presentation highlighted how investigation of the exposome has become tractable, as well as the important roles that individuals can play in supporting this effort.

In the following interview, Dr. Tatonetti discusses some of the approaches his team is using to explore the exposome.

Air pollution in Beijing, China. Photo by Kentaro Iemoto, Wikimedia Commons.

What do you see as the key differences between what you are doing with "big data" and earlier scientific methods?

In some sense, computational approaches like the ones my lab uses are actually just a high-dimensional approach to the way science has always worked. Years ago Darwin sailed to the Galapagos Islands, and he drew a beautiful figure of a tree in his notebook that represents the relationship he observed between the beaks of finches and their geographic distribution on the islands. He was able to contain his observations and interpretation of the data on a single piece of paper. Today, every genome sequencing run generates a terabyte of data, and medical records contain petabytes worth of data. We can't possibly hold all of the variables in these kinds of data matrices in our heads anymore, and yet we know that there must be valuable insights hidden in there somewhere.

What we're trying to do is to bring the technology of digesting observations and producing good scientific hypotheses up to speed with our ability to generate and collect these data. Instead of just looking at a little bit of data and coming up with one hypothesis, we process terabytes and petabytes of data, generate thousands of high-confidence hypotheses, and then evaluate and validate them just like we would any other scientific hypothesis. If a hypothesis turns out to be true we follow it up with other kinds of computational and experimental studies. If we find evidence against it we throw it out and go to the next one.

How does your own research fit into this framework?

Although my early research received some attention for identifying adverse drug events and drug-drug interactions, the thread that holds my work together is my interest in coming up with new ways of analyzing observational data sets; that is, large collections of data that are gathered opportunistically. For example, when Google records search queries and results, it accumulates large numbers of observations, providing unique opportunities to objectively identify trends within the data.

The problem is that data sets generated in this way present challenges for analysis. For example, imagine that you saw a large uptick in searches for sexually transmitted diseases on a particular day. Does this mean that all of the people doing those searches contracted an STD? This is what Google Flu would assume. It's actually more likely that there was a big news story on the topic that day that brought STD's to people's attention. A simple thought experiment like this suggests that making inferences from observational data can be incredibly tricky, because events happen in parallel with the data you are collecting, and you can't measure what you don't capture. My work focuses on finding better ways to analyze such observational data sets without jumping to the wrong conclusions.

If factors that are important to the data analysis aren't actually contained in the data, how do you compensate for them?

In the example I just gave there is just one variable — the number of searches for STD's per day. There's no way to account for hidden information in a case like that. But if you can collect hundreds of different variables or, in the case of electronic health records, tens of thousands of different variables, you have lots of dimensions to explore.

The general strategy we've been using begins with the fact that a large data set offers many more dimensions to explore than you can actually use in your analysis. If you are interested in knowing whether a drug is correlated to an outcome or adverse drug reaction, you're essentially using two variables out of a 10,000 variable data set. Our hypothesis is that those other 9,998 variables can tell you something about the underlying structure of the data.

To untangle the data we build a covariance matrix of all possible variables, and then look at variables that uniquely identify a patient population. For example, we might segregate the data set into whether a patient was exposed to a specific drug or not. We can then use this classification to identify good controls for the population. Then, we look for another subset in the data that is similarly structured and compare it to the control. In doing so, we assume that the biases inherent in the data — for example, missing information or confounding variables — are going to align. So we don't correct for the bias, but look for another equally biased sample. As long as the biases align, the difference between the two populations should be the effect that we want to measure.

What does all of this have to do with the exposome?

Just like the genome is the complete collection of all of our genes, or the transcriptome is all of the RNAs that have been transcribed in the cell, the exposome is the complete collection of everything a person has been exposed to during his or her life. It could include things like pollutants in air or water, chemicals from eating fish or meats, effects of living in different environments around the world, or prescription drugs, just to name a few. It's another kind of very large, high-dimensional data set.

My collaborator, Chirag Patel, who graduated from Stanford the same year I did, has been looking at factors in the environment

that affect health. He famously conducted a study he called an EWAS — an environment-wide association study. This was the first time that every possible environmental factor collected by the National Health and Nutrition Examination Survey was correlated with all possible diseases. He found some very interesting results, focusing primarily on toxins, which are often structured like small molecule drugs. Meanwhile, I was working on characterizing interactions between small molecule drugs and the body. Fundamentally we're actually working on the same problem, and so this collaboration seemed like a perfect fit.

The exposome could theoretically include pretty much anything, and so it presents a similar problem as the STD example I mentioned earlier. Our approach is to focus on exposures that we can measure and quantify. For example, even when sufficient data aren't available for the concentration of specific toxins in the environment, quantitative measurements characterizing overall air quality often exist. As long as we can quantify something, it goes into the exposome, creating an observational data set we can then interrogate using computational methods.

With this approach, is it possible to connect exposures to genetic traits within individuals that might make them more or less susceptible to a particular risk factor?

It's still very early, but we expect that there will be certain responses to environmental factors that are shared across all humans, and others that are unique to specific individuals. Not a lot is known about how the body interacts with the environment. Clearly, it does a lot to keep itself healthy in the presence of all types of pollutants and toxins. At the same time, though, there are some people who are missing a component that's necessary to maintain health. This missing component might only become apparent when a person is exposed to a particular toxin, chemical, or drug and gets sick. At the same time, a person with a normal genotype might be perfectly fine in that polluted environment or when treated with that same drug.

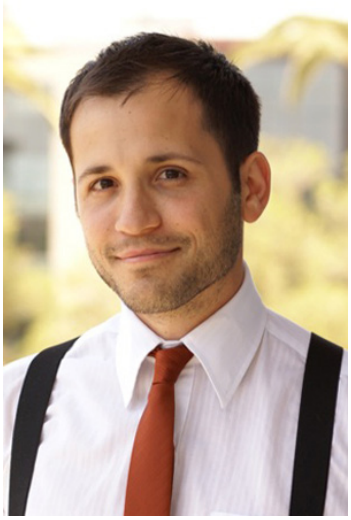
In the past, epidemiological studies have identified exposures, like smoking, that lead to changes in health, like lung cancer or diabetes. What's new in the exposome project is that we are developing data science tools that consider all environmental factors at once. This should help us to holistically understand how interactions among multiple factors influence individual responses to drugs and the environment.

What kinds of data do you use to construct your data sets?

In thinking about exposures like pollution in the environment, we can use databases developed by the Environmental Protection Agency that quantify levels of known toxins across the whole country, sometimes even at the county or city level. We can then integrate these data with health records from different locations and map their differential exposures down to the zip code. With these pieces in place, we can look at what diseases correlate to those exposures across geographical areas, highlighting potential interactions with environmental factors.

In my lab we are trying to collect enough data about the drugs, the exposures, and the diseases, so that we can form hypotheses about which conditions might be connected with genetic causes. We recently began collaborating with David Goldstein using Columbia's medical records system, where we are identifying cohorts with unique diseases or drug responses that we suspect may be the result of an underlying genetic factor. We may then sequence and run proteomic analysis of these patients to identify previously unknown genetic variants in human disease.

How can the public at large help in contributing to our understanding of the exposome?



Electronic health records are obviously one important data source, but new personal health monitoring technologies like the Apple Watch or the Fitbit could also potentially have a role to play in this. One could imagine an app, perhaps enabled by Apple's HealthKit, that allows patients to participate in research trials, lets them explore their long-term health statistics, and compare to their peers.

The goal would be to enable a critical mass of users who are actively collecting data about themselves, creating another kind of observational data set. This would allow us to spontaneously spawn research studies in response to new hypotheses. Ideally, we would like to improve on the current system where a finding goes to publication, then someone gets an idea for a way to implement it, then public health agencies pick it up. Currently, it just takes too long for a discovery to have an impact, and we would like to change that.

Columbia Investigators Awarded New NCI Physical Sciences in Oncology Center

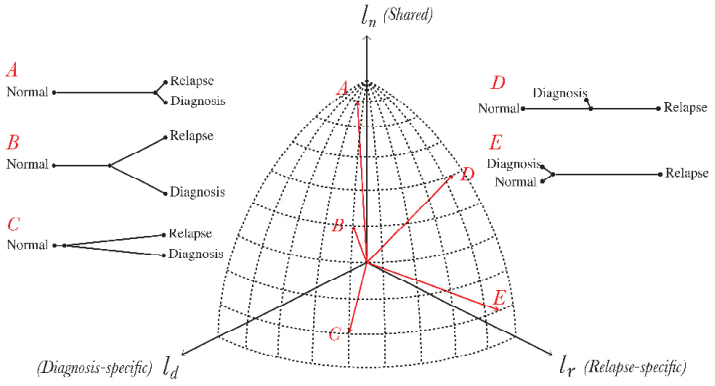
The National Cancer Institute's Physical Sciences in Oncology program has announced the creation of a new center for research and education based at Columbia University. The Center for Topology of Cancer Evolution and Heterogeneity will develop innovative mathematical and experimental techniques to explore how genetic diversity emerges in the cells that make up solid tumors. In this way it will address a key challenge facing cancer research in the age of precision medicine — how to identify the clonal variants that are responsible for a tumor's growth, spread, and resistance to therapy.

Co-directing the Center are Columbia University's Raul Rabadan and Antonio Iavarone. The multidisciplinary research team also includes 8 additional investigators. Other participating scientists from the Department of Systems Biology include Michael Shen, Peter Sims, Chris Wiggins, and Hossein Khiabanian.

In addition to conducting research, the Center will create the New York Metropolitan Area Discussion Group in Mathematics and Oncology, which will promote dialogue between physical scientists and mathematicians on the one side, and cancer biologists on the other. The new Center will also provide opportunities for quantitative investigators to become embedded in biology laboratories; distribute startup grants for interdisciplinary, collaborative research; and organize an annual international conference.

The key scientific issue driving research at the new center is an emerging awareness that genetic diversity among the cells is a critical factor in determining how a tumor grows. The Center will address this issue by developing new interdisciplinary methods for understanding how such diversity develops. Focusing on prostate cancer and glioblastoma as sample systems, it will develop widely applicable pipelines for modeling tumor evolution, dissecting clonal heterogeneity, and achieving insights into the genetic mechanisms that lead to drug resistance during cancer treatment.

Work at the Center will make use of two key, emerging experimental technologies. One is the organoid culture, in which three-dimensional buds of organs are grown in the laboratory, providing an experimental system that more closely replicates what happens in the body than traditional *in vitro* cultures. As Center investigator Michael Shen has shown, these can be combined with fluorescent tagging methods that make it possible to track how specific clones divide and evolve during tumor development. A second key technology is high-throughput single-cell sequencing. New techniques currently under development in the laboratory of Peter Sims make it possible to quickly generate readouts of the genomes and gene products of large numbers of individual cells. These experimental methods will enable the Center to provide a high-resolution representation of genetic differences among the cells that make up a tumor, making it possible to distinguish how clonal subpopulations evolve.



Investigators at the new Center for Topology of Cancer Evolution and Heterogeneity are investigating how mathematics could inform the analysis of cancer genomic data.

Another critical element of the Center's approach will be the development of new mathematical approaches for interpreting the large, high-dimensional data sets that such single-cell technologies generate. In recent work, Raul Rabadan, Gunnar Carlsson, Andrew Blumberg, and other participating scientists have been developing new applications of a mathematical field called topological data analysis (TDA), an advanced method for characterizing the relationships and structures underlying large collections of discrete measurements.

Using TDA, they will create a mathematical structure for identifying distinct cell populations, and determining their association with the progression of primary solid tumors. Using another mathematical concept called moduli spaces, they will also define methods for comparing "trees" representing clonal evolution from multiple clones. Finally, they will design algorithms to disentangle relationships among mutational events in tumor evolution, to distinguish mutations that truly drive cancer progression from those that are merely associated with it.

By combining experimental and quantitative approaches in this way, the Columbia University Center for Topology of Cancer Evolution and Heterogeneity will deliver validated methods for inferring clonal evolution, new single-cell genomic protocols for uncovering clonal heterogeneity, and experimentally validated machine learning approaches for predicting drug sensitivities. "Our ultimate goal," Dr. Rabadan says, "will be to provide the cancer research community with a framework for unraveling complexity in solid tumors, with the long-term aim of improving diagnosis and treatment."

For more information about the Center for Topology of Cancer Evolution and Heterogeneity, visit **http://psoc.c2b2.columbia.edu**.

Tracing Bacterial Evolution Across Billions of Years

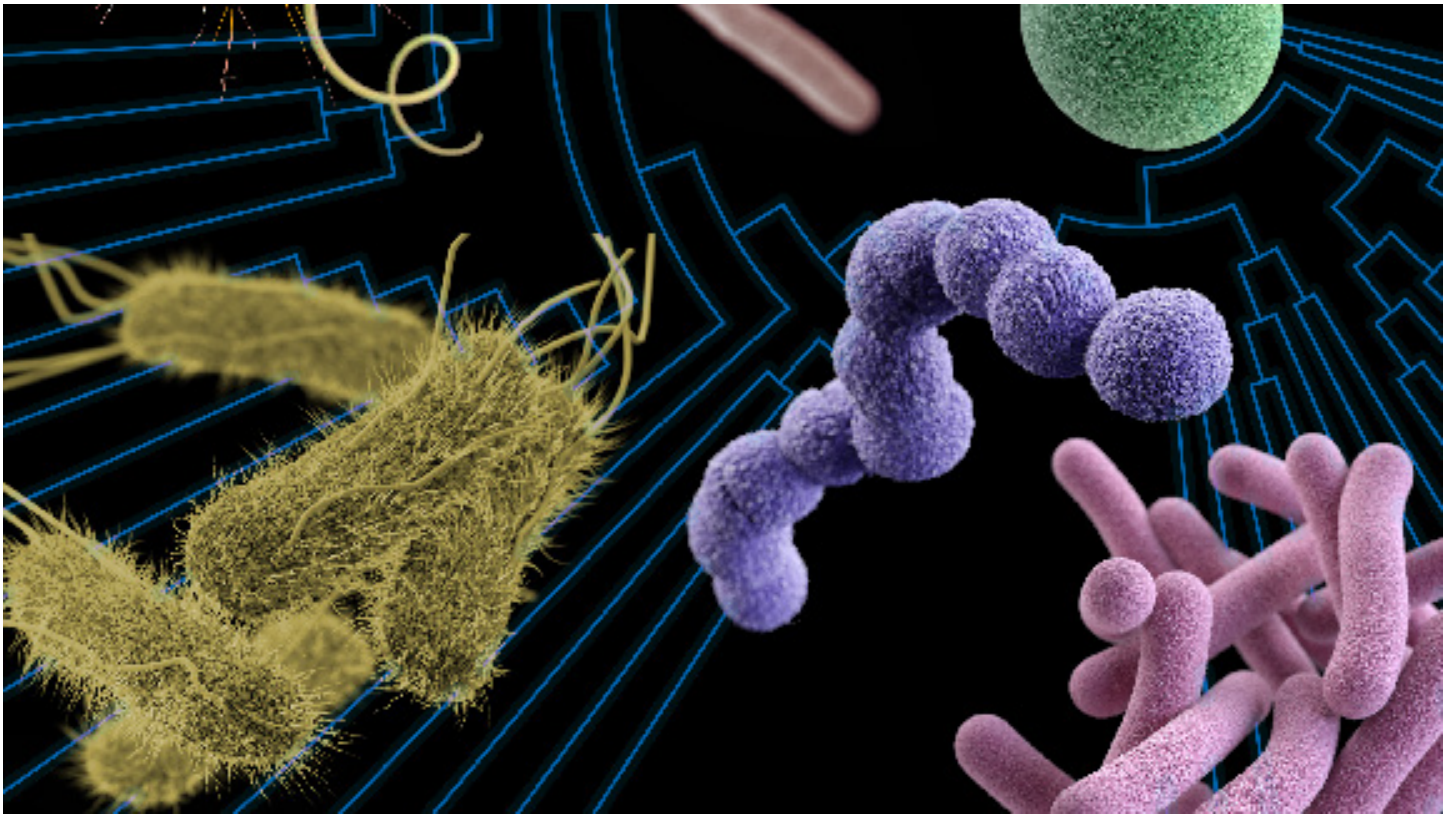


Image courtesy of Germán Plata and Dennis Vitkup. Article reprinted with permission from *Columbia News*.

Bacteria might not get the same prominent placement in the Museum of Natural History as, say, those dinosaurs on the fourth floor. They might not have their own plastic toys in the gift shop. In the arena of evolutionary staying power, however, these little guys are real champions.

Bacteria appeared on Earth at the dawn of life itself, about 3.5 billion years ago. Since that time they've displayed amazing feats of evolutionary adaptation, living high in the stratosphere and many miles below the surface of Earth, in cold arctic lakes and superhot thermal vents.

But these tiny creatures are not simply passive adapters or opportunists. They played a major role in shaping our environment. Microbes created the atmospheric oxygen we breathe and are now responsible for replenishing Earth's ecosystems by degrading waste products and recycling important nutrients.

Bacteria, argues Dennis Vitkup, an associate professor at Columbia's Departments of Systems Biology and Biomedical Informatics, also have an intimate relationship with every one of us. "If you consider the total number of cells within your body, there are about 10 times more bacteria than human cells. We lead a very synergistic life with them," Vitkup says. The contribution of bacteria to our well-being and environment is usually invisible, obscured from us due to their size. But invisible does not mean unimportant. "Life on Earth as we know it would quickly stop," says Germán Plata, a postdoctoral researcher on the Vitkup team, "if bacteria were to suddenly disappear."

Despite their omnipresence, microbial evolutionary adaptations are often challenging to study, partly due to the difficulty of growing diverse bacteria in the lab. "Probably less than a dozen bacteria are really well studied in the laboratory," Vitkup says.

Writing in the journal *Nature* this past January, Vitkup and Plata applied computational tools to investigate bacterial evolutionary adaptations by simulating metabolism for more than

Related publication:
Plata G, Henry CS, Vitkup D. **Long-term phenotypic evolution of bacteria**. Nature. 2015 Jan 15;517(7534):369-72.

300 bacterial species, covering the entire microbial tree of life.

To make sense of the puzzling nature of evolutionary adaptations over billions of years, Vitkup invokes Darwin's famous finches. Over the course of a couple million years, finches on different islands developed different forms of beaks that allowed them to pick up and eat different kinds of seeds. The differences in animal appearance, such as beak shapes or even behavior, are the external manifestations of genetic differences, what researchers usually call phenotypic differences. "But how do you compare phenotypic differences among bacteria over billions of years?" Vitkup asks. Vitkup and Plata started with one key idea: that the crucial factor in bacterial survival is their ability to grow on different food sources. A particular bacterium's phenotype depends on its ability to convert available nutrients into the chemicals that it needs to grow and reproduce. The team then used bacterial metabolic simulations to predict phenotypes across hundreds of environments. "Bacterial metabolism serves as a kind of microscope to discern patterns of phenotypic adaptation and diversification," Plata says. Vitkup's team discovered that long-term bacterial adaptation proceeds via two different stages. Initially, there is a relatively fast diversification lasting tens of millions of years, which is then followed by a slow divergence process continuing for billions of years.

Surprisingly, the team was able to describe the average long-term diversification by a simple mathematical model. It shows that, on average, there is a constant rate of phenotypic change on the scale of billions of years, revealing the continuity of bacterial adaptation since the beginning of life on our planet.

Vitkup, a native of Ukraine, studied theoretical physics at the Moscow Institute of Physics and Technology. In 1992, at age 21, he moved to the U.S. to work on a Ph.D. in biophysics at Brandeis University and ended up working in the Harvard lab of Martin Karplus (the 2013 Nobel Prize winner in chemistry), using computers to simulate protein dynamics.

Vitkup arrived at Columbia in 2004 after post-doctoral fellowships at MIT and Harvard Medical School. Plata, who is Vitkup's former student, graduated with a degree in biology from Colombia's National University. He joined the graduate program here in 2008 after working as a plant biotechnology researcher in his native country.

For the last decade, the Vitkup group has been developing computer algorithms that use DNA sequences and what's known about biochemical interactions to spit out a list of metabolic reactions likely present in a given microbe. These methods could be used to fight deadly pathogens. Indeed, the algorithms have already been used to piece together the metabolism of a malaria parasite. "It allowed us to predict about 40 new therapeutic targets," Plata says. Last year Vitkup obtained a grant from the National Institutes of Health to build accurate metabolic models for all major human bacterial pathogens. The computational algorithms developed in the Vitkup lab make it possible to analyze the hundreds of thousands of bacterial genomes which will be sequenced over the next decade. "As with simulations in the field of nuclear research," Vitkup says, "we are moving toward an era when most of this work will be done on a computer and then, in some cases, results will be experimentally investigated in the lab." The next challenge for Vitkup and colleagues is stepping up the degree of complexity and understanding how hundreds of bacteria work together in symbiotic microbial communities. It is exciting, says Vitkup, because "we now have the tools to really understand the beauty and complexity of the invisible microbial world."

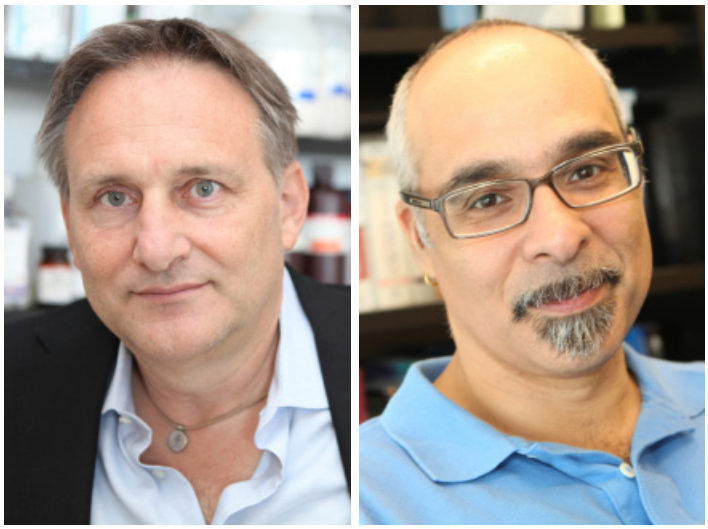
Rodney Rothstein Elected to National Academy of Sciences

The Columbia University Department of Systems Biology congratulates Rodney Rothstein on his election to the National Academy of Sciences. The NAS is a private, non-profit society of distinguished scholars that provides independent, objective advice to the nation on matters related to science and technology. Scientists elected to the NAS are chosen by their peers in recognition of their distinguished and continuing achievements in original research.

Dr. Rothstein is professor of genetics & development at Columbia University Medical Center and holds an interdisciplinary appointment in the Department of Systems Biology. He has pioneered the use of recombination to alter genomes and has used these methods to isolate novel genes involved in the maintenance of genome stability. His development of "one-step" gene disruption technology led directly to the "knockout" technology used in many organisms to exploit recombination to either remove or insert DNA sequences into specific positions within the genome.

Dr. Rothstein is the second member of the Department of Systems Biology to be nominated to the National Academy of Sciences. Barry Honig, director of the Columbia University Center for Computational Biology and Bioinformatics became a member in 2004. Also among this year's NAS inductees is Ricardo Dalla-Favera, a collaborator in the Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet).

Using Master Regulators to Reclassify Cancer Subtypes



Andrea Califano and Aris Floratos

Andrea Califano and Aris Floratos, faculty members in the Columbia University Department of Systems Biology, have received a two-year, $624,236 subcontract to develop a new classification system of cancer subtypes. The agreement was awarded through a subcontract from Leidos Biomedical Research, Inc., which operates the Frederick National Laboratory for Cancer Research for the federal government.

By performing an integrative analysis of genomic data from the Cancer Genome Atlas (TCGA) and proteomic data from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC), the researchers plan to recategorize tumors collected in TCGA based on the master regulator genes that determine their state. This is in contrast to other approaches based on expression of genes that reflect tissue lineage and proliferative processes. In addition, the team will link the genetics of each tumor sample to the specific master regulators that determine its state using a recently published novel algorithm (DIGGIT). Ultimately, the project aims to provide a more useful catalog of pan-cancer subtypes that could help to identify biomarkers and therapeutic targets for specific kinds of tumors, and ultimately provide a resource to guide the next generation of precision medicine.

"We have to reevaluate the way in which we organize tumors within subtypes, using both gene expression data and mutational data," says Dr. Califano. "Right now the common approach is to classify tumor types based on rather generic genes that are differentially expressed between subtypes. But most of these genes play no role in actually driving the disease. We want to shift the emphasis and classify tumors based on the genes that truly regulate tumor state and survival."

Over the past 10 years, the Califano Lab has developed a suite of computational methods for modeling cell regulatory interaction networks (also called interactomes), demonstrating that interactomes of particular cancer subtypes become "rewired" in consistent and predictable ways. They have also repeatedly found that such networks allow systematic identification of genes called master regulators, which represent regulatory bottlenecks that are necessary and sufficient to establish and maintain tumor state. Such master regulators are infrequently mutated and thus evade detection by conventional mutational studies. Yet a number of them are essential for tumor survival and many occur in synthetic lethal pairs, where neither gene in isolation is essential for tumor survival but the pair is.

Under this new subcontract, the Califano Lab will apply this perspective to identify the master regulators of every tumor represented in TCGA, on a sample-by-sample basis. In addition to using TCGA data — which contains data about a tumor's mutations, gene expression, and other genomic information — they will also make use of proteomic data from the CPTAC Data Portal, which contains mass spectrometry measurements related to protein identity, protein abundance, and post-translational modifications that can degrade a protein or change its regulatory activity. Incorporating such high-quality experimental data into this systems biology-based computational approach will make it possible to develop reliable models of how the various proteins in the network work together to drive disease.

Once the master regulators for all of the tumors in the database have been identified, tumor samples in the top 20 cancer types represented in TCGA will be reclassified, using a pan-cancer approach. Doing so, Califano anticipates, will reveal a limited repertoire of master regulators that ultimately drive a large faction of the tumors, many of which should be independent of traditional organ based tumor classification. Using the DIGGIT algorithm, they will also look upstream of master regulators, within regulatory networks, to identify the genomic and epigenomic alterations that determine their aberrant activity.

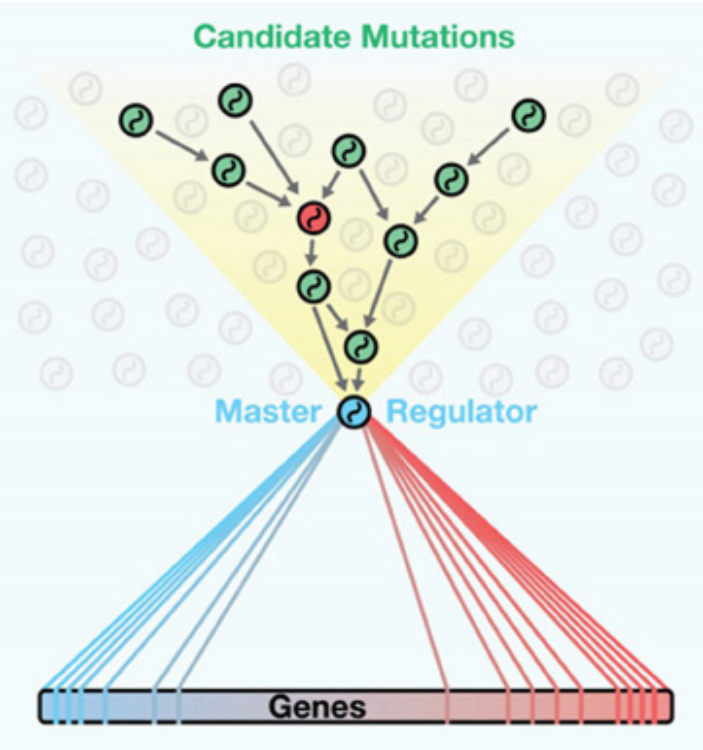
Knowing the mutations and master regulators that drive these newly defined cancer subtypes will dramatically simplify the landscape for future cancer genome research. The project is just beginning, but based on initial data Califano anticipates that they might find that as few as 250 proteins, out of a genome containing more than 20,000 genes, may turn out to be responsible for driving a majority of tumors.

"Based on findings we published in a study of prostate cancer in 2013," Califano says, "we think this limited number of master regulators will be essential in selecting a very useful panel of secreted protein biomarkers that could be monitored in the blood. Once they are identified, our hope is that a test could be devel-

oped that would look for those proteins or DNA transcripts in a blood test, and provide valuable information that could be used to guide personalized early cancer detection and treatment."

As a final step, the Califano Lab will search for existing, FDA-approved drugs that have already been shown to target the mutations and master regulators that their analyses reveal. Applying their findings in this way could also yield practical insights that could be used to control or eliminate tumors. These will complement traditional genetic-based approaches that match small molecule inhibitors to mutated oncogenes.

To maximize the impact of the project's findings, Dr. Floratos will oversee the integration of all data into geWorkbench, a web-based portal that provides easy access to the Columbia University Department of Systems Biology's computational tools. In addition, the entire computational pipeline developed for the project, called Citrus, will be implemented as a cloud-based service on Google's Compute Engine Infrastructure. This will enable cancer researchers anywhere in the world to access the data and regulatory models that the project generates, and to perform their own analyses as new cancer data become available.



The DIGGIT algorithm looks upstream of master regulators to distinguish driving mutations of cancer from passenger mutations.

N-of-1 Clinical Trials for Cancer Started

In 2014, a team led by Andrea Califano, in collaboration with Gary Schwartz, Edward Gelmann, and other physicians at Columbia University Medical Center (CUMC), launched an innovative new approach to clinical trials aimed at improving precision medicine for cancer. This methodology, called an N-of-1 clinical trial, uses the master regulator model of cancer developed in the Califano Lab to identify critical bottlenecks in the genetic networks that drive individual patients' tumors.

CUMC is enrolling 260 patients to participate in N-of-1 clinical trials. Tumor types currently under investigation include gastrointestinal neuroendocrine tumors, glioma, gastric adenocarcinoma, lung adenocarcinoma, pancreatic adenocarcinoma, soft tissue sarcoma, metastatic or triple negative breast carcinoma, and squamous cell carcinoma of the head and neck.

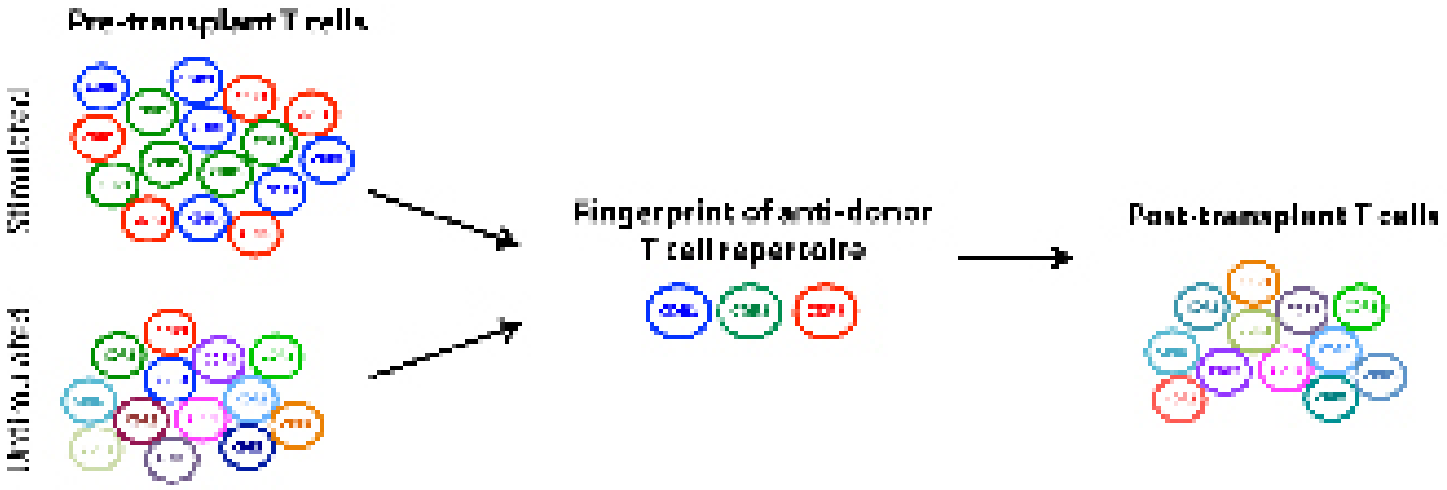
During the N-of-1 trial, tumor tissue removed from patients undergoes RNA expression analysis. The resulting data are then analyzed within the context of models of gene regulation developed in the Califano Lab that are specific to that tumor type. This makes it possible to identify the distinctive master regulators of cancerous activity in individual patients' tumors.

The researchers then search for FDA-approved drugs — or drugs currently at advanced stages of clinical testing — that are known to target those master regulators. Suitable drugs are immediately tested on the patient's tumor sample, either in cell culture or after being implanted into a mouse model (patient-derived xenograft), to determine if the compound stops tumor growth. If treatment is effective, it may be investigated further in a more traditional clinical trial.

Past research has provided strong indications that this systems biology-based approach could offer a powerful new paradigm for cancer diagnosis and for identifying more effective, precision treatment strategies. This is still an experimental method, however, and so these N-of-1 clinical trials are necessary to test its effectiveness.

The scientists expect that master regulators identified in one tumor may also play important roles in others. As the results of N-of-1 studies accumulate, investigators will begin to assemble a more precise, more comprehensive understanding of different cancer subtypes. In this way, N-of-1 clinical trials should help to focus scientists' attention on the best therapeutic opportunities, accelerating future cancer research.

Mechanism of Kidney Transplant Tolerance Discovered



After identifying T cell clones that react against donated kidney tissue in vitro, new computational methods developed in Yufeng Shen's Lab are used to track their frequency following organ transplant. The findings can help to predict transplant rejection or tolerance.

When a patient receives a kidney transplant, a battle often ensues. In many cases, the recipient's immune system identifies the transplanted kidney as a foreign invader and mounts an aggressive T cell response to eliminate it. To minimize complications, many transplant recipients receive drugs that suppress the immune response. These have their own consequences, however, as they can lead to increased risk of infections. For these reasons, scientists have been working to gain a better understanding of the biological mechanisms that determine transplant tolerance and rejection.

Yufeng Shen, an assistant professor in the Columbia University Department of Systems Biology and JP Sulzberger Columbia Genome Center, together with Megan Sykes, director of the Columbia Center for Translational Immunology, recently took an encouraging step toward this goal. In a paper published in *Science Translational Medicine*, they report that the deletion of specific donor-resistant T cell clones in the transplant recipient can support tolerance of a new kidney.

Enabling their specialized function within the immune system, all T cells are defined by their T-cell receptor, a protein located on their surface that recognizes "foreign" agents from outside the body. As T cells mature, a somatic process called V(D)J recombination produces diverse populations of clones. In aggregate, these clones present a broad spectrum of T cell receptors, enabling the immune system to respond to many different pathogens. "It's like having a standing army," Shen explains. "You want a wide variety of T cell receptors even before a pathogen enters the body so that you have soldiers ready to respond to it."

This constant state of preparedness becomes a problem, however, following an organ transplant, as a subset of T cell clones reacts much as it would to infection. To understand this mechanism, the investigators wanted to determine which clones are responsible. To do so, they utilized an *ex vivo* assay called mixed lymphocyte reaction (MLR). In this experiment, blood from the recipient and a small amount of donor tissue are placed together in a tube before the transplant. When brought together in this way, lymphocytes — a class of immune cells that includes T cells — react to the tissue in the same way one would expect they would in the body. The investigators then perform bulk deep sequencing on the sample to identify T cell clones that react against the donor tissue. This is possible because the genetically diverse coding regions for the T cell receptor serve as a kind of "barcode" for each clone sequence.

In the research described in the paper, kidney transplants were then performed on six patients. Two utilized conventional transplant methods, while four used an experimental technique that Dr. Sykes developed called combined kidney and bone marrow transplantation (CKBMT). In this procedure, the kidney transplant is accompanied by the transplantation of bone marrow from the donor to the recipient. Once incorporated into the recipient's body, the donated bone marrow begins generating its own T cells, leading to the development of a hybrid immune system that can recognize the donated organ as "self." Following kidney transplant in all six patients, blood was drawn at 6, 12, and 18 months, and additional deep sequencing analyses were performed at each time point. This enabled the investigators to determine how the donor-reactive T cell clones changed due to interactions between the recipient's immune system and the donated kidney.

The behavior of the donor-reactive T cell clones in these follow-up se-

quencing experiments provides clues about how well the recipient's immune system tolerates the foreign kidney. If there is an expansion in the frequency of donor-reactive clones following transplantation, the researchers hypothesized that there should be a reaction against the donated organ, usually indicating transplant rejection. In the study, this phenomenon occurred in the two patients who received a conventional kidney transplant without bone marrow transplant. Intriguingly, the investigators found that in 3 of the 4 patients receiving CKBMT, the frequency of donor-reactive clones identified in the original MLR experiment decreased, and each patient tolerated the transplant. In the fourth patient, the number of donor-reactive cells did not decrease, and the patient rejected the kidney.

These observations suggest that the new hybrid immune system can initiate what the researchers characterize as a "deletion mechanism" that actively destroys or suppresses the expansion of donor-reactive T cells. Although this surprising finding is still under investigation, it suggests a mechanism behind transplant tolerance, and provides exciting evidence that CKBMT could offer a strategy for improving the success of kidney transplantation. Currently, the investigators are planning a clinical trial of this method.

Essential to this promising discovery was a new statistical method the Shen Lab developed for analyzing T cell genome sequencing data. "Although the audience for the paper is mostly immunologists and transplant experts," he remarks, "more than two-thirds of the time spent on this project focused on analyzing data. We had to develop an entirely new approach for analyzing T cell diversity from scratch."

Because there are limits to the amount of blood that can be drawn from a patient, it is extremely difficult using standard genome analysis methods to distinguish rare T cell clones. This is because clones that appear at low frequencies become indistinguishable from statistical noise caused by inherent and unavoidable limits in the accuracy of genome sequencing technologies. Although having a larger sample of blood cells could theoretically improve statistical power, it is only clinically feasible to sample a relatively small portion of the billions of T cells and hundreds of millions of T cell clones that circulate in the body. "It's like trying to infer properties of the ocean by taking just a drop of water," Shen says.

Borrowing ideas from statistical physics and information theory, however, he and graduate student Boris Grinshpun devised a set of methods to quantify diversity and divergence of T cell samples. Specifically, their analysis made use of statistical methods including the Simpson diversity index and generalized entropy (GE) index, techniques for measuring the concentrations of individuals when classified into types. By tracking how frequently different clonal states occur and how their frequencies change over time, the investigators can observe how the immune system is responding to the transplant. If there is a very active reaction to a donor's tissue, the diversity of the T cell population will decrease, with high numbers of clones appearing at high frequency. If the recipient tolerates the transplant, the diversity will be maintained at a normal level, with limited amounts of expansion in the number of clones.

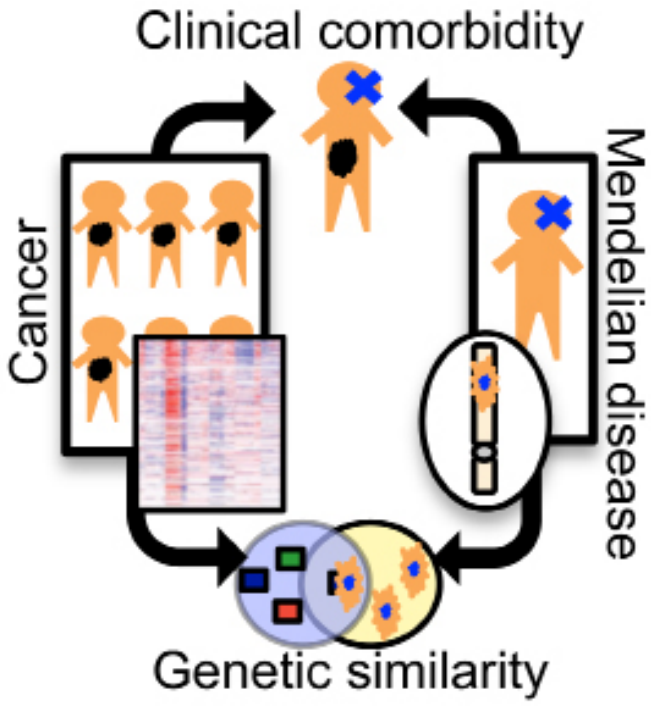
Yufeng Shen

By using bulk sequencing to look at the ensemble properties of the T-cell clone distribution in this way, and comparing them across multiple time points, Shen's method provides a general method for tracking the dynamic properties of the recipient–transplant interaction. Identifying clonal expansion in this way could potentially help predict transplant rejection before it occurs, giving physicians evidence to support the administration of more aggressive immunotherapy. And in cases in which the analysis points to transplant tolerance, this approach could indicate when treatment with immunosuppressive drugs could be safely scaled back.

The approach used in this paper is indicative of the Shen Lab's general interest in developing new ways to accurately identify rare clonal variants. In ongoing projects, his lab is working with the Columbia Genome Center to optimize new experimental approaches to sequence T cell receptor genes and develop related computational methods for extracting signal from noisy data sets. These approaches could have applications not just in organ transplantation, but also in cancer immunotherapy and studies of autoimmune diseases.

Related Publication:
Morris H, DeWolf S, Robins H, et al. **Tracking donor-reactive T cells: evidence for clonal deletion in tolerant kidney transplant patients**. Sci Transl Med. 2015 Jan 28;7(272):272ra10.

Connections Found Between Mendelian Diseases and Cancer



Comorbidity between Mendelian diseases and cancer may result from shared genetic factors.

Mendelian disorders occur when specific mutations in single genes — called germline mutations — are inherited from either of one's two parents. Well-known examples of Mendelian diseases include cystic fibrosis, sickle cell disease, and Duchenne muscular dystrophy. Other genetic diseases, including cancer, result from somatic mutations, which occur in individual cells during a person's lifetime. Because the genetic origins of Mendelian diseases and cancer are so different, they are typically understood to be distinct phenomena. However, scientists in the Columbia University Department of Systems Biology have found evidence that there might be interesting genetic connections between them.

In a paper published in *Nature Communications*, postdoctoral research scientist Rachel Melamed and colleagues in the laboratory of Raul Rabadan report on a new method that uses knowledge about Mendelian diseases to suggest mutations involved in cancer. The study takes advantage of an enormous collection of electronic health records representing over 110 million patients. The authors show that comorbidity of Mendelian diseases and cancer can be tied to genetic changes that play roles in both diseases. The paper also identifies several specific relationships between Mendelian diseases and the cancers melanoma and glioblastoma. The article grew from Dr. Melamed's PhD thesis, completed earlier this year, which built on a previous collaboration between the Rabadan Lab and that of Andrey Rzhetsky (University of Chicago). In that earlier work, the investigators mined a huge collection of electronic health records to determine the extent to which genes responsible for Mendelian diseases also raised a patient's risk of developing other complex diseases. In her thesis, Melamed ran with this idea in the context of cancer genomics, hypothesizing that comorbidity between a Mendelian disorder and cancer might result from overlap between the genes responsible for each disease.

She and her colleagues first analyzed data from the same collection of electronic health records used in the earlier study, looking for patients who both had a Mendelian disorder and developed cancer. They then investigated whether the genes responsible for the Mendelian disorders were related to genes frequently mutated across more than 5000 tumors represented in The Cancer Genome Atlas. Using multiple measures of genetic similarity — including shared genes, shared molecular pathways, and gene–protein interactions — they showed that there is greater than expected similarity between the genes involved in Mendelian disease and the somatic alterations involved in the comorbid cancers seen in those patients.

The authors also discovered several specific connections between Mendelian disorders and cancer. For example, they found that genes associated with melanoma are also found in oculocutaneous albinism, a disorder that affects the pigmentation of the skin, hair, and eyes. In addition, investigations of Diamond-Blackfan anemia, a blood disorder, and holoprosencephaly, a disorder in cranial development, identified genetic connections with glioblastoma, the most aggressive form of brain cancer.

Summarizing the implications of these discoveries, the authors write, "[T]his suggests that comorbidity between Mendelian disease and cancer may be due to germline mutations that provide a fertile ground for the growth of certain aberrant cells." They also write that some Mendelian diseases predispose patients to multiple types of cancer, a finding that is in line with a growing body of evidence that many tumor mutations have "pan-cancer" effects.

The Rabadan Lab's method is interesting not only because it provides a clearer picture of the genetic basis of comorbidities, but also because the cancer-associated genes it identifies are likely to be present in some tumors even in the absence of comorbid Mendelian disease. It therefore complements other methods for identifying cancer genes.

Related publication:
Melamed RD, Emmett KJ, Madubata C, et al. **Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes**. Nat Commun. 2015 Apr 30.

Novel Machine Learning Method Expands Landscape of Breast Cancer Driver Genes

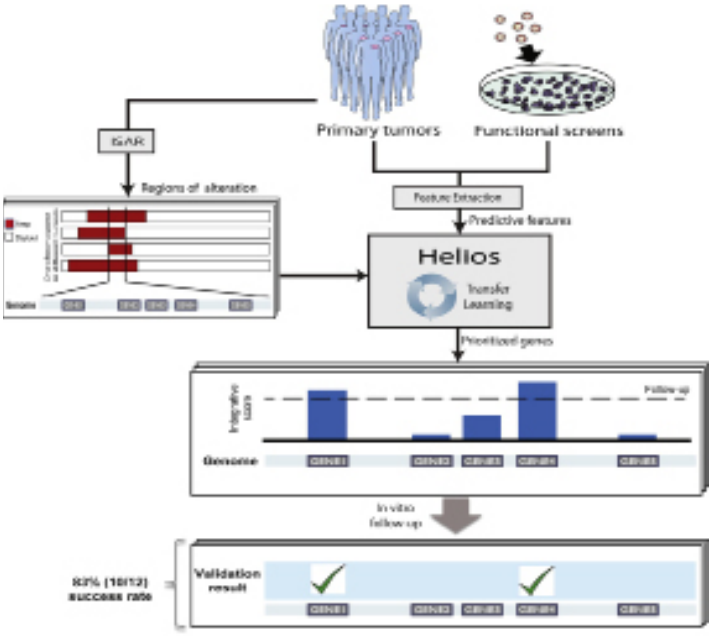
For many years, researchers have known that somatic copy number alterations (SCNA's) — insertions, deletions, duplications, and transpositions of sections of DNA that are not inherited but occur after birth — play important roles in causing many types of cancer. Indeed, most recurrent drivers of epithelial tumors are copy number alterations, with some found in up to 40% of patients with specific tumor types. However, because SCNA's occur when entire sections of chromosomes become damaged, biologists have had difficulty developing effective methods for distinguishing genes within SCNA's that actually drive cancer from those genes that might lie near a driver but do not themselves cause disease.

In a paper published in *Cell*, researchers in the laboratories of Dana Pe'er and Jose Silva (Icahn School of Medicine at Mount Sinai) report on a new computational algorithm that promises to dramatically improve researchers' ability to identify cancer-driving genes within potentially large SCNA's. The algorithm, called Helios, was used to analyze a combination of genomic data and information generated by functional RNAi screens, enabling them to predict several dozen new SCNA drivers of breast cancer. In follow-up *in vitro* experimental studies, they tested 12 of these predictions, 10 of which were validated. Their findings nearly double the number of known breast cancer drivers.

The method reported in the paper incorporates two algorithms. The first, called Identification of Significantly Altered Regions (ISAR), improves upon previous algorithms for identifying SCNA's expected to harbor a driver gene by accounting for variations in the local rate at which copy number alterations occur, due to features such as DNA secondary structure and epigenetic alterations. When the researchers applied ISAR to 785 breast cancer samples, they identified 83 significantly amplified SCNA-containing regions, more than doubling the 30 regions previously reported in TCGA.

A second algorithm, called Helios, integrates additional information — in this case point mutation, gene expression, and functional RNAi screening data — into a single candidate driver score. By using machine learning techniques to identify complementary patterns within these diverse data types, Helios iteratively prioritizes genes within significantly altered regions that have the highest probability of being true cancer drivers. As lead author Felix Sanchez-Garcia explains, "Other people used copy number variation to try to narrow things down to a single gene, which is impossible for many regions. Instead, we use copy number to guide interpretation of all the other features."

Using this integrative approach, Helios correctly scored 13 out of 14 (93%) drivers ranked highest within significantly amplified regions. In addition, 10 out of 12 genes (83%) that Helios predicted to be cancer drivers were validated in experimental studies. Reflecting on these results, Dr. Pe'er explains, "This is the first and largest scale systematic validation undertaken for an algorithm of this type, and its accuracy is unprecedented. Its ability to reveal so many new cancer drivers is due to the fact that our hypotheses were generated in an unsupervised way using statistical criteria, rather than by cherry-picking our candidates based on prior biological knowledge. The experimental results show that Helios is a very robust algorithm that can generate biological insights that would be extremely difficult to produce in any other way."



Helios produces an integrated score by combining features derived from primary tumors and genome-wide shRNA screens.

The findings dramatically expand the landscape for investigating drivers of breast cancer, and offer enormous potential for translational breast cancer research. Copy number alterations are known to be much more common than point mutations in patients with breast cancer, which means that this paper's findings offer the opportunity to identify diagnostic and therapeutic strategies that could improve treatment for large numbers of women with the disease. Moreover, although used in this paper to study breast cancer, Helios is capable of identifying somatic copy number alterations in many other cancer types in which SCNA's play a role.

Related publication:
Sanchez-Garcia F, Villagrasa P, Matsui J, et al. **Integration of genomic data enables selective discovery of breast cancer drivers**. Cell. 2014 Dec 4;159(6):1461-75.

Grants, Awards, and Honors

Cory Abate-Shen and **Michael Shen** received an NIH/NCI R01 award for their project titled "Investigating the cell of origin for bladder cancer."

Juan Arriaga, a postdoc in the Abate-Shen Lab, received a Prostate Cancer Research Program Postdoctoral Training Award from the Department of Defense (DoD) office of the Congressionally Directed Medical Research Programs.

Aditya Dutta, an Associate Research Scientist in the Abate-Shen Lab, is one of 16 winners of the Irving Institute/Clinical Trials Office Pilot Awards. His project is titled "Identification of Molecular Drivers of Cancer Cell Adaptation to Metabolic Stress."

Adolfo Ferrando received a three-year award from the Pershing Square Sohn Cancer Research Alliance for "Functional Dissection of Oncogenic Enhancers in T-Cell Leukemia."

Oliver Hobert, a three-year grant from the National Institute of Mental Health for "Developing Drivers for Neuron Type-Specific Gene Expression." He was also one of 401 investigators selected to the Fellows of the American Association for the Advancement of Science.

Anupama Khare, a postdoctoral scientist in the Tavazoie Lab, has received a K99 Pathway to Independence Award from the National Institute for Allergy and Infectious Diseases for a project titled "Dissection of Complex Multifactorial Interactions between Bacterial Species."

Richard Mann received a grant from the National Institute of Neurological Disorders and Stroke for a project titled "Functional Mapping of Pathways for Sensory-Motor Integration." He also received a three-year award from the National Institute of General Medical Sciences for "Proximo-Distal Patterning in the Drosophila Appendages" in a competitive renewal.

Maho Shibata, a postdoctoral research scientist in Michael Shen's lab, received an NIH/NCI K99 award for her project "Investigation of luminal stem cells and castration resistance in prostate cancer."

Columbia was selected as one of 12 institutions nationwide to receive a Beckman Scholar Award to help recruit and train the most outstanding undergraduate students in biology and chemistry. The governance committee for the award will be chaired by **Brent Stockwell**.

Roxanne Toivanen, a postdoctoral research scientist in Michael Shen's lab, received an Australian National Health and Medical Research Council (NHMRC) CJ Martin Biomedical Early Career Fellowship entitled "Systems analyses of prostate cancer organoid cultures for precision medicine."

Harris Wang was named a winner of an Alfred P. Sloan Research Fellowship and received an award from the National Science Foundation for "A Systems Approach to Study Horizontal Acquisition of Regulatory DNA."

Sakellarios Zairis, a graduate student in the Rabadan and Wiggins Labs, received a TL1 Personalized Medicine Training Award from the Irving Institute for Clinical and Translational Research. His project is titled "Genomic Characterization of NASH-Related Hepatocellular Carcinoma."

PhD Graduates

Congratulations to our recent graduates from Department of Systems Biology laboratories.

Miki Hayano (Stockwell Lab)
Anna Kaplan (Stockwell Lab)
Rachel Melamed (Rabadan Lab)
Kenichi Shimada (Stockwell Lab)
Vasanthi Viswanathan (Stockwell Lab)
Jung Hoon Woo (Califano Lab)

COLUMBIA SYSTEMS BIOLOGY

Contact Us

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

Christopher M. Williams
Communications Director
Department of Systems Biology
cmw2189@cumc.columbia.edu

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