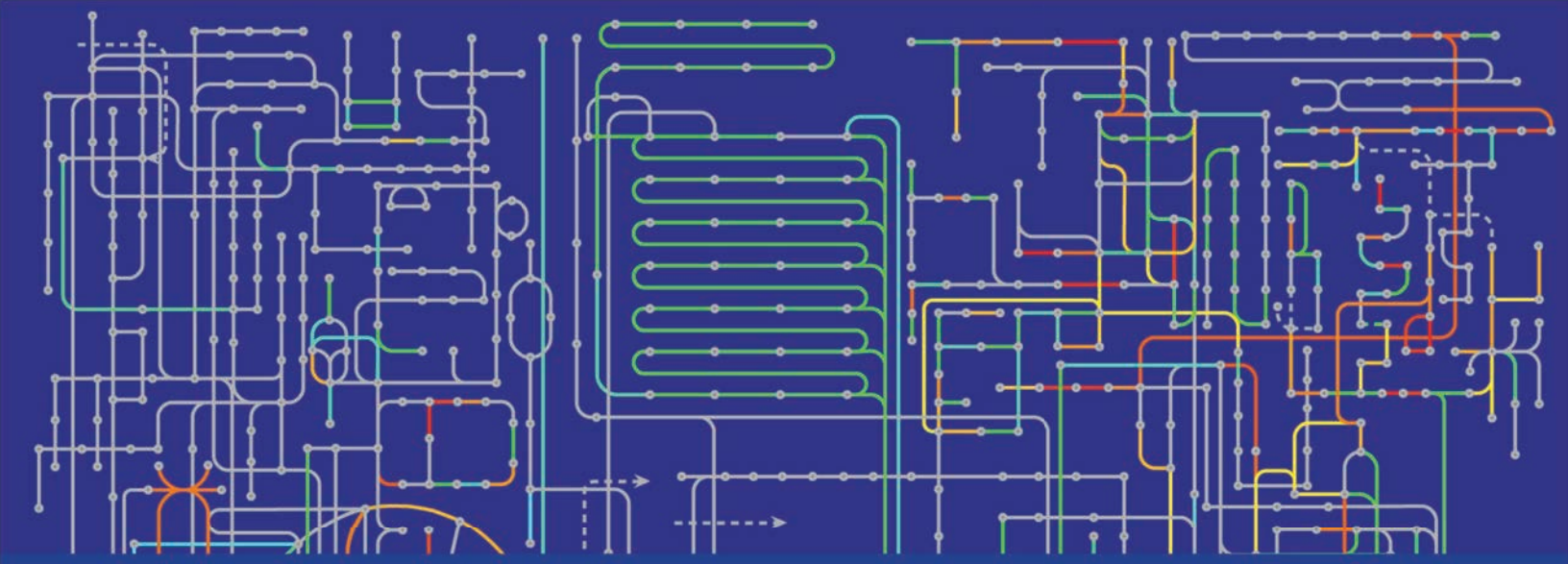


Columbia University
Department of Systems Biology

Research Highlights, 2014

COLUMBIA UNIVERSITY
MEDICAL CENTER

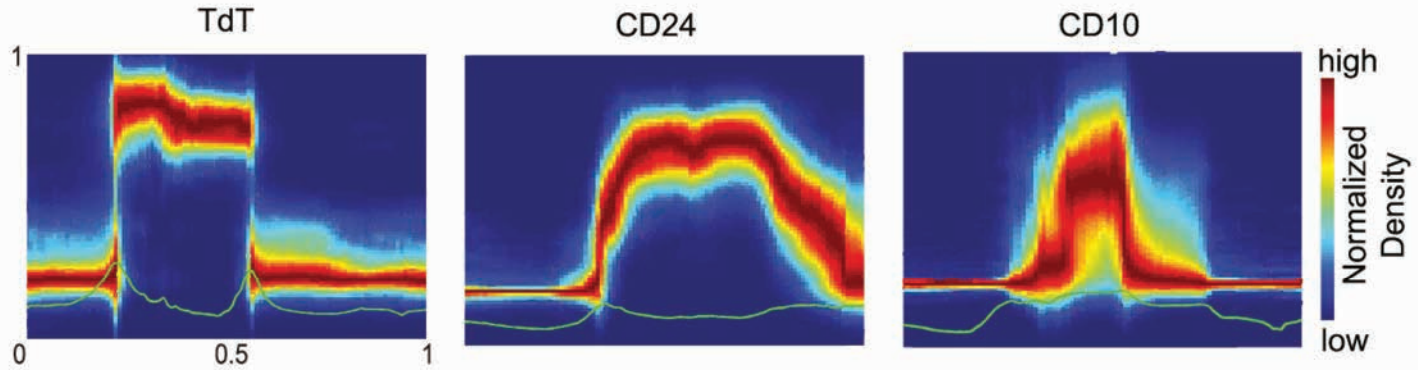
A Comprehensive Map of Human B Cell Development

In a paper published in the journal *Cell*, a team of researchers led by Dana Pe'er at Columbia University and Garry Nolan at Stanford University describes a powerful new method for mapping cellular development at the single cell level. By combining emerging technologies for studying single cells with a new, advanced computational algorithm, they have designed a novel approach for mapping development and created the most comprehensive map ever made of human B cell development. Their approach will greatly improve researchers' ability to investigate development in cells of all types, make it possible to identify rare aberrations in development that lead to disease, and ultimately help to guide the next generation of research in regenerative medicine.

Pointing out why being able to generate these maps is an important advance, Dr. Pe'er, an associate professor in the Columbia University Department of Systems Biology and Department of Biological Sciences, explains, "There are so many diseases that result from malfunctions in the molecular programs that control the development of our cell repertoire and so many rare, yet important, regulatory cell types that we have yet to discover. We can only truly understand what goes wrong in these diseases if we have a complete map of the progression in normal development. Such maps will also act as a compass for regenerative medicine, because it's very difficult to grow something if you don't know how it develops in nature. For the first time, our method makes it possible to build a high-resolution map, at the single cell level, that can guide these kinds of research."

Although conventional dogma characterizes development as a series of discrete steps, cells actually develop in a process of continuous transformation. As a cell matures, the complex and interconnected molecular programs that regulate its activity change gradually. These programs can also differ significantly among cells of similar type, and in many important regulatory cell types, rare aberrations in these programs can have devastating effects. Previously, it was difficult to observe these subtle differences in sufficient detail to distinguish them, but new technologies now offer important new opportunities. Just as genome sequencing transformed how biology was studied in the previous decade, new technologies for analyzing the molecular properties of single cells are currently revolutionizing the kinds of questions many biologists are asking. Dr. Pe'er sees single-cell approaches as an important step beyond genomics. "DNA sequencing can identify genes and mutations, but often they are not studied in context," she points out. "With single-cell approaches, we can map the cells where the action actually happens and what the genes are doing inside them. Single-cell mapping will do for development what genome sequencing has done for genetics."

In the research described in the *Cell* paper, the investigators used an emerging technology called mass cytometry, which in a single experiment can measure 44 molecular markers simultaneously in millions of individual cells. This provides a wealth of data that can be used to compare, categorize, and chronologically order cells, and makes it possible to begin identifying the molecular systems responsible for development in much greater detail than was ever before possible.



A new algorithm called Wanderlust uses single-cell measurements to detect how marker expression changes across development.

Taking advantage of this complex single-cell data also required the researchers to develop new mathematical and computational methods for interpreting it. Just as one can represent a physical object in three dimensions — length, width, and height — the Pe'er lab's approach involved thinking of the 44 measurements as a 44-dimensional geometric object. They then developed a new computational algorithm, called Wanderlust, which uses mathematical concepts from a field called graph theory to reduce this high-dimensional data into a simple form that is easier to interpret.

Such a high-dimensional geometry is impossible for us to visualize, and so Wanderlust converts the collection of measurements of important developmental markers in each cell into a single one-dimensional value that corresponds to the cell's place within the chronol-

Related publication:

Bendall SC, Davis KL, Amir ED, et al. **Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development**. Cell. 2014, Apr 24;157(3):714-25.

ogy of development. By including all of the hundreds of thousands of cells that were profiled, this graph does not force cells into categories, but represents a continuous new geometry that captures the entire developmental trajectory. The researchers can then use these trajectories as a scaffold for characterizing the order and timing of molecular and regulatory events during the developmental process and understand how the process unfolds in health and disease.

According to Dr. Pe'er, "For years, mathematics has been used in physics to show that there are very elegant mathematical relationships that define the way the universe is structured. When looked at from the perspective of mathematics, we see that the complexity found in biology is also beautifully structured and patterned. Our body has trillions of cells of countless different types, each type bearing different molecular features and behavior. This complexity expands from a single cell in a carefully regulated process called development. This regulation creates patterns and shapes in the high-dimensional data we measure. By using Wanderlust to analyze these data we can find the pattern and trace the trajectory that cellular development follows."

To test their approach, the researchers studied development in human B cells, a type of cell that is important in the adaptive immune response and is involved in a variety of autoimmune diseases as well as certain types of cancer. Since immune cells continue to develop throughout adult life, even a single sample from one bone marrow contains cells from all stages of B-cell development. Investigators in the Nolan lab used mass cytometry to profile 44 markers in a cohort of approximately 200,000 healthy immune cells that were gathered from one such sample. In each cell they measured cell surface markers that help identify what type of cell it is, as well as markers inside the cell that can reveal what the cell is doing, including markers for signaling, the cell cycle, apoptosis, and genome rearrangement.

Using Wanderlust to analyze the high-dimensional data provided by mass cytometry, the researchers accurately ordered the entire trajectory of 200,000 cells according to their developmental chronology. In a strong indication of its accuracy, Wanderlust captured and correctly ordered all of the primary molecular landmarks known to be present in human B cell development. Wanderlust also pinpointed a number of previously unknown key regulatory signaling checkpoints that are required for human B cell development as well as uncharacterized subtypes of B cell progenitor cells that correspond to important developmental stages. This trajectory constitutes the most comprehensive analysis of human B cell development that has ever been conducted.

In a sign of the high degree of precision that Wanderlust offers for studying single cells, the researchers also report that they identified rare, previously unknown signaling events involving the signaling protein STAT5 that occurred in just 7 out of 10,000 cells. This regulatory event is involved in the process of VDJ recombination, a volatile time when the B-cell is reshuffling its own DNA. Further laboratory experiments showed that disrupting these signaling events using kinase inhibitors fully stalled the development of B cells.

Identifying and characterizing the regulatory checkpoints that control and monitor cell fate can have many practical applications, as many diseases result from imbalance in the cell types produced by the immune system. Thus, the approach that the authors describe can produce insights that could be used for the development of new diagnostics and therapeutics. This process for mapping how healthy cells develop can be applied not just to B cells, but to any type of cell. As the authors suggest, their method offers the possibility of providing a foundation for studying normal development as well as the processes responsible for any kind of developmental disease. In the future, they anticipate that more sophisticated models could be built that would be capable of representing even more complex systems, such as the entire immune system. The holy grail would involve creating a complete map of every cell type in the body and how each progressively develops from a single stem cell, pinpointing every cell fate decision along the way.

Achieving this goal will require a lot more work, but as Dr. Pe'er points out, "This current project is a landmark both in the study of development and in single-cell research, and has completely changed the way I think about science. A fire has been lit, and these findings are just the tip of the iceberg of what is now possible."

Dana Pe'er Wins 2014 Overton Prize and NIH Pioneer Award

Dana Pe'er was named one of 10 winners of the 2014 NIH Director's Pioneer Awards, which support "high risk, high reward" science that holds great potential to transform biomedical or behavioral research. Dr. Pe'er's award will support an ambitious new project to develop the technological and computational methods necessary to create a complete, cell-type by cell-type atlas of how all cells in the human body develop from a single embryonic cell.

"Just as the human genome project mapped all the genes in our body and became a basis for studying genetic variation," she explained, "the next grand challenge is to catalog all cell types in our body and organize them based on their lineages. Having this reference would provide an essential foundation for many areas of research. The Pioneer Award will provide vital funding that will help us to achieve this goal."

Dr. Pe'er was also named the winner of the 2014 Overton Prize. Awarded by the International Society for Computational Biology (ISCB), the prize recognizes one outstanding scientist each year who has made a significant contribution to the field of computational biology. The award recognizes Dr. Pe'er "for her cutting-edge research that applies computational methods to complex data to understand the organization of molecular networks in cells at a holistic systems level."

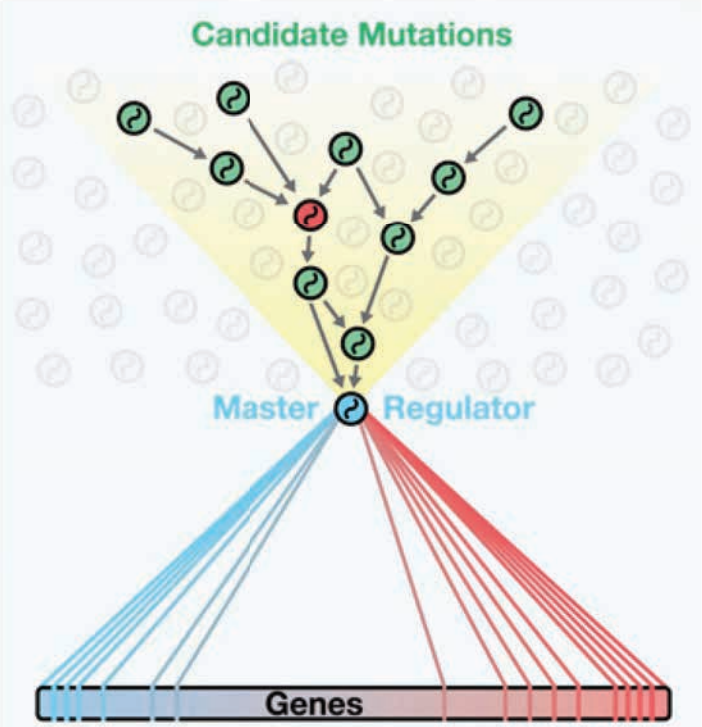
Algorithm Identifies Genetic Driver of Mesenchchymal Glioblastoma

Although genome-wide association studies have made it possible to identify mutations that are linked to diseases such as cancer, determining which mutations actually drive disease and the mechanics of how they do so has been an ongoing challenge. In a paper published by *Cell*, researchers in the lab of Andrea Califano describe a new computational approach that may help address this problem.

The manuscript presents an innovative algorithm called DIGGIT (short for Driver-Gene Inference by Genetic-Genomic Information Theory), which utilizes and adds an important new method to the toolbox the Califano Lab has been assembling over the past 10 years for modeling and interrogating regulatory networks. The fundamental idea behind it is that any mutation that represents a causal driver of a disease must be upstream of the "master regulator" genes that are the functional drivers the disease. (Master regulators are genes whose activity, either alone or in synergy with other genes, is essential for the onset and persistence of diseases such as cancer.) By thinking of master regulators as a "funnel" through which the pathways connecting the mutation to the disease phenotype are obligated to pass, DIGGIT can systematically identify the relatively small number of mutated genes that could drive disease.

Using this approach, the team discovered that loss of the gene *KLHL9* drives the mesenchymal subtype of glioblastoma (GBM), the most deadly type of brain tumor. They also validated this finding through extensive experiments and in additional cohorts. Specifically, the team found that tumor growth in mesenchymal GBM lacking *KLHL9* can be suppressed by reintroducing the gene. DIGGIT also accurately identified disease drivers in breast cancer and Alzheimer's disease, suggesting that the algorithm, combined with MARINa, the Califano Lab's extensively published tool for identifying master regulators of complex disease, now offers a powerful new method for identifying genetic drivers of other types of disease as well.

"We see this paper as a culmination of our efforts," Dr. Califano says, "because it shows that in diseases as diverse as Alzheimer's and different types of cancers, there are regulatory bottlenecks that integrate upstream signals, and channel them into the aberrant activity of master regulators. These, in turn, activate the pathological programs that are necessary for the emergence of disease."



A new algorithm called DIGGIT identifies mutations that lie upstream of crucial bottlenecks within regulatory networks. These bottlenecks, called master regulators, integrate these mutations and become essential functional drivers of diseases such as cancer.

How DIGGIT works

The genesis of DIGGIT began with James Chen, a recent PhD graduate and member of the Califano Lab, as well as the first author of the paper. In the mesenchymal glioblastoma study, Chen began with gene expression and mutational profile data of more than 250 patients collected by the Cancer Genome Atlas consortium. The first step in the algorithm is to perform a genome-wide analysis to eliminate any genomic copy number variations (CNVs) that are clearly incapable of perturbing the molecular network of the tumor. DIGGIT filters the data to retain only CNVs at loci that are predicted to have cis-regulatory effects (changing gene expression of the genes in which they are located). This important first step eliminates CNVs that cannot cause changes in gene expression in the tumors, thus drastically amplifying the statistical power of any given collection of patient samples. In this way it overcomes a limitation of traditional genome-wide studies, which require large cohorts or large effect sizes to produce statistically robust results.

The remaining "functional" variants, or fCNVs, are then checked to see if they are predictive of increased activity of the master regulators connected with disease. DIGGIT evaluates the fCNVs using ssMARINA (single sample Master Regulator Inference Algorithm, another tool developed in the Califano Lab), which provides a quantitative measure of how active each master regulator is in each patient sample in the cohort. fCNVs that appear to alter master regulator activity are retained and then evaluated using MINDy (Modulator Inference by Network Dynamics), which predicts genes that are upstream of master regulators within the genetic network. Any fCNVs that are associated with a change in gene expression but are not connected to a master regulator in the interaction network are then eliminated from consideration. In this way, the combination of algorithms acts as a computational sieve, specifically capturing those genes that

affect the activity of master regulators of a specific disease trait.

Finally, the best candidate driver mutations are identified using an analysis derived from classical genetic tests. The approach was developed to address the confounding issue that genomic alterations in cancer rarely occur individually. Rather, deletions and duplications take place across sections of chromosomes, leading to statistical dependencies between mutations; that is, genes that are next to each other are far more likely to be mutated together than genes that are more distant from each other. This also implies that any mutation that drives disease will be accompanied by additional mutations that are associated with disease only because of their proximity to the driver. To gain a clear picture of the genetic causes of disease, it is critical to distinguish the true driver mutations and artifacts of the analysis.

The team hypothesized that for any set of genes that are associated with a phenotype, no artifact gene can be more associated with that phenotype than the driver mutation. Using the list of candidate genes identified in the previous steps, the algorithm computationally assessed every gene in turn for its association with the mesenchymal GBM subtype. Through this process of elimination, DIGGIT narrowed the list of possible genes into a small list of loci that it predicted to be essential for the mesenchymal phenotype.

KLHL9 deletion drives mesenchymal GBM

The DIGGIT pipeline identified two candidate genes that could be responsible for driving mesenchymal glioblastoma. The first, *C/EBP-δ*, had previously been identified in a collaboration with Antonio Iavarone, a professor of pathology and cell biology and neurology in the Columbia University Institute for Cancer Genetics, as playing a role in the development of mesenchymal tumors, so Chen focused on the second, which codes for a protein called KLHL9 that had never been identified as being relevant in brain cancer.

A series of follow-up laboratory experiments demonstrated that loss of *KLHL9* leads to aberrant activity in the mesenchymal master regulators *C/EBP-β* and *C/EBP-δ*. When normal *KLHL9* function is restored, it suppresses the activity of these two genes by mediating the degradation of the proteins. Further tests explored the effects of *KLHL9 in vivo* by implanting mesenchymal glioblastoma cells into living mice. In mice in which *KLHL9* expression was restored, tumor growth was significantly impaired. Taken together with results of several additional experiments described in the paper, the findings reveal that deleting both copies of *KLHL9* is sufficient to transform GBM tumors to an aggressive, mesenchymal subtype. Moreover, rescuing normal expression of *KLHL9* is sufficient to severely hamper tumorigenesis of mesenchymal GBM. Considering that 50% of people with mesenchymal glioblastoma exhibit *KLHL9* mutations, the findings suggest a potentially valuable therapeutic strategy for assessing brain cancer, by concentrating on the bottleneck that integrates the mutations as opposed to focusing on the mutations themselves.

Additional applications of DIGGIT

While implemented and validated mechanistically in GBM, DIGGIT can be used to investigate any phenotype for which matched gene expression and variant profiles (either somatic or germline) are available for a sufficient number of samples. To show that the algorithm could be used to identify genetic drivers of other diseases, Chen conducted additional analyses looking at *BRCA*-positive breast cancer and Alzheimer's disease.

In the breast cancer study, he first searched the scientific literature for copy number variants already linked to breast cancer. He uncovered 25 alterations that had previously been reported as being associated with this breast cancer subtype. He then performed the DIGGIT analysis, comparing gene expression in breast cancer cells to that of normal breast tissue cells as controls. The algorithm identified 35 genes as drivers of *BRCA* breast cancer. Of the 25 genes previously identified in the literature, 19 (76%) appeared in the DIGGIT analysis, suggesting that the algorithm is highly capable of capturing driver mutations in other cancer types. It also revealed a number of never-before-seen genes that may warrant further investigation.

In the study of Alzheimer's disease, DIGGIT identified 14 statistically significant variants that appear to drive the condition. Among these, the highest ranked was a variant in the gene *TYROBP*, which researchers at the Icahn School of Medicine at Mount Sinai independently predicted to be a driver of late-onset disease for the first time in 2013. DIGGIT also identified the *APOE* locus, a well-known variant associated with Alzheimer's disease.

Because DIGGIT identifies mutations within the context of genome-wide regulatory networks, Dr. Chen points out that it offers an important advantage over traditional gene association methods. "Not only does looking at CNVs through the lens of master regulators dramatically increase our ability to detect candidates even in highly heterogeneous populations," he explains, "it also provides a direct mechanistic perspective on exactly how these genes initiate the disease. There are many instances in which the identification of a candidate gene via genome-wide association studies precedes elucidation of its mechanism by years. Even in our studies of breast cancer and Alzheimer's disease, where the goal was simply to show that DIGGIT could identify candidates that are missed by more traditional methods, it provided the additional benefit of immediately identifying the key molecular regulators and pathways that the mutations likely work through to produce a disease."

Now a joint postdoctoral scientist in the laboratories of Angela Christiano and Andrea Califano, Dr. Chen continues work to incorporate the algorithm into Bioconductor, a widely used bioinformatics platform. In the meantime, the Califano Lab has begun to incorporate DIGGIT into its pipeline for integrating genetics and genomics across all tumors, including 20 in the Cancer Genome Atlas.

Related publication:
Chen JC, Alvarez MJ, Talos F, et al. **Systematic analysis of regulatory networks reveals KLHL0 as a novel genetic determinant of the mesenchymal subtype of glioblastoma**. Cell. 2014 Oct 9;159(2):402-14.

Diverse Autism Mutations Lead to Different Disease Outcomes

People with autism have a wide range of symptoms, with no two people sharing the exact type and severity of behaviors. In a large-scale analysis of hundreds of patients and nearly 1000 genes, a team of researchers led by Associate Professor Dennis Vitkup has started to uncover how diversity among traits can be traced to differences in patients' genetic mutations.

Autism researchers have identified hundreds of genes that, when mutated, likely increase the risk of developing autism spectrum disorder (ASD). Much of the variability among people with ASD is thought to stem from the diversity of underlying genetic changes, including the specific genes mutated and the severity of the mutation. "If we can understand how different mutations lead to different features of ASD, we may be able to use patients' genetic profiles to develop accurate diagnostic and prognostic tools and perhaps personalize treatment," Dr. Vitkup said.

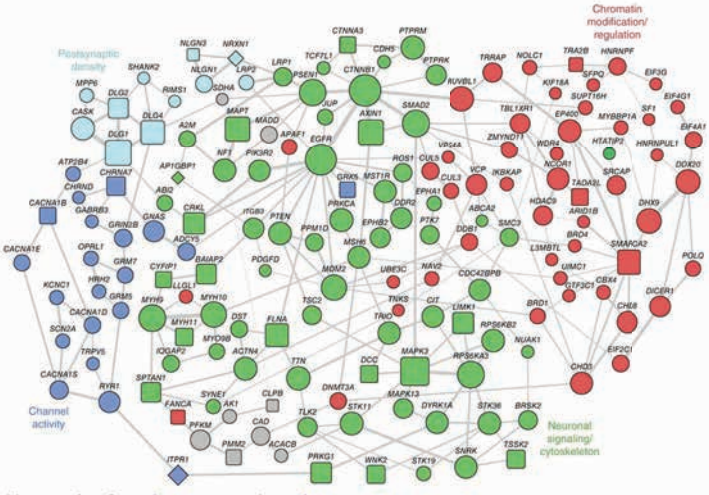
IQ and gender differences in autism influenced by severity of mutations

To investigate the links between genetic mutations and autism traits, Dr. Vitkup and a team of Columbia graduate students analyzed genetic and clinical data on hundreds of patients with ASD from the Simons Simplex Collection. The group found that more damaging genetic mutations usually lead to worse disease outcomes. "It looks as if high-IQ autism cases are usually triggered by milder mutations," Dr. Vitkup said.

Patients with low-verbal or nonverbal IQs usually had mutations in genes that are more active in the brain. And high-IQ individuals were less likely to have mutations that completely shut down genes. Instead, mutations that only partially damage normal gene function in the brain appear to be predominantly associated with high-functioning autism cases.

Gender differences in autism could also be traced to the types of genes mutated in the individual. Though ASD is far more common in males, females with ASD are more likely to fall on the severe end of the spectrum. The Columbia researchers found that the genes mutated in females generally had greater activity throughout the brain than those mutated in males. Very damaging ASD mutations in girls on average are found in genes that are almost twice as active as typical genes in normal brains.

"These patterns are consistent with the idea that there are mechanisms that protect females," Dr. Vitkup said. "Most often, only when a mutation hits a highly active gene do we see symptoms in females. Given that the inherent differences in gene activity in male and female brains are typically on the order of a few percent, these findings are quite remarkable."



Network of autism-associated genes.

Autism mutations disrupt multiple cell types

Behavioral variability in autism patients may also stem from the types of brain cells affected, and the Columbia researchers have taken the first steps in determining which cell types in the brain are most affected by autism mutations. The team identified these cells by looking at the normal activity of autism-related genes in dozens of similar cell types in mouse brains. The analysis showed that many different types of neurons throughout the brain are affected by mutations in autism genes. "The idea that eventually all autism mutations would converge onto a single type of neuron or single brain area isn't what we see in the data," Dr. Vitkup said. "Instead, an autism mutation usually affects multiple brain areas simultaneously."

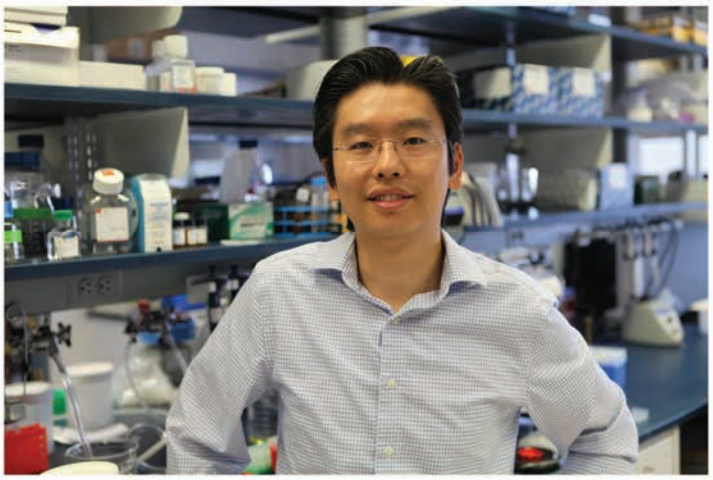
Certain neurons, however, appear to be more affected than others. The Columbia researchers found strong effects in cortical and striatal neurons that form a circuit that controls repetitive motions and behaviors, such as rocking, an insistence on sameness, and restricted interests, which are common in people with ASD. "There are many hypotheses about the types of neurons and circuits involved in autism, but by using unbiased genome-wide approaches, like the one used in this study, one can understand which neurons are the most important and explain the core features we see in people with ASD," said Dr. Vitkup.

Identifying the circuits involved is the next step in understanding autism, he said. "Huge progress has been made in the last five years: We and our colleagues have now identified multiple affected genes, and we are coming to a consensus about how the genes work together in biological networks. Now, based on the affected genes, we are identifying affected cell types and brain circuits and trying to connect them to disease outcomes in individual patients."

Related publication:
Chang J, Gilman SR, Chiang AH, Sanders SJ, Vitkup D. **Genotype to phenotype relationships in autism spectrum disorders**. Nat Neurosci. 2014 Dec 22.

New Directions in Genome Engineering: An Interview with Harris Wang

As a graduate student in George Church's lab at Harvard University, Harris Wang developed MAGE, a revolutionary tool for the field of synthetic biology that made it possible to introduce genomic mutations into E. coli cells in a highly specific and targeted way. Now an Assistant Professor in the Columbia University Department of Systems Biology, Dr. Wang recently published a paper in ACS Synthetic Biology that introduces an important advance in the MAGE technology. The new technique, called (MO)-MAGE, uses microarrays to engineer pools of oligonucleotides that, once amplified and integrated into a genome, can generate thousands or even millions of highly controlled mutations simultaneously. This new method offers a cost-effective way for designing and producing large numbers of genomic variants and provides an efficient platform for experimentally exploring genome-wide landscapes of mutations in bacteria and optimizing the organisms' biochemical capabilities.



In the following interview, Dr. Wang explains the origins of the new technology, and discusses what he sees as the remarkable potential it holds for both basic biological research and industrial applications of synthetic biology.

How are MAGE and (MO)-MAGE different from more traditional methods in genome engineering?

In traditional genome engineering, researchers would induce genome perturbations randomly. For example, you might use ultraviolet radiation or a mutagen to generate mutations and then do a selection experiment to compare and isolate cells with different genotypes based on how they respond to specific stimuli. The problem with this approach, though, is that you have no way to control what mutations occur, even if you know the mutation you are interested in investigating.

In the late 1990s and early 2000s this led people to begin thinking about how to produce mutations in a targeted way. One important step forward occurred when Don Court and Barry Wanner independently developed a homologous recombination system using lambda red proteins. Their method enables double stranded recombination at a specific location much more efficiently than was possible beforehand. When I say "more efficient," though, the efficiency was still only something like 10-4 (just 1 in 10,000 cells), and the technique only made it possible to induce a single mutation at a time.

Things started to move forward when Court discovered in the early 2000s that small oligonucleotides of 70-80 base pairs could be incorporated into the genome at very high efficiency by targeting them to specific sites of interest in homology arms. As a graduate student in George Church's lab, I used this concept to design a way of doing targeted genome engineering across many different positions in the *E. coli* cell. You would design an oligonucleotide that, at either end, had a set of base pairs that was complementary to the genetic site of interest and also included the mutation you wanted to insert. If you knock out the genetic pathway through which cells usually repair replication errors, these oligos can then bind with their target sites during replication and are incorporated into the genome approximately 25% of the time. By repeating this process multiple times, you could then quickly increase the percentage of mutated cells in the population. We called this technology MAGE, which stands for multiplexed automated genomic engineering.

How did (MO)-MAGE grow out of this work?

After we developed MAGE, our next question was how to make it more cost-effective, which is critical because engineering a biological pathway can often require mutating a large number of genes. If you wanted to target thousands of sites simultaneously, you would need to order thousands of oligos, which is far beyond the financial resources of any lab.

We thought of a few potential ways to solve this problem, but the most tantalizing one was based on synthesis from microarrays. Although microarray synthesis has typically been limited to fewer than 60 base pairs, Agilent and other companies have recently developed the capability to synthesize up to 230 base pairs at high fidelity. This offers a great opportunity, though when you receive these oligos on a microarray, they come in just picoMolar concentrations. Because MAGE requires large numbers of oligos, (MO)-MAGE (which stands for microarray oligonucleotide-MAGE) borrows methods from gene synthesis to amplify the numbers of oligos a million-fold. Using common amplification primers, we design and synthesize a pool of reactions that contains sublibraries within the pool.

In the (MO)-MAGE paper, for example, we describe a chip we designed that simultaneously introduces 13,000 mutations. Each of these features was grouped into one of 8 different classes, each of which had a different function. For example, one class included oligos designed to knock out the open reading frames in *E. coli* by introducing a stop codon and then frameshift mutation.

Another class inserted a T7 polymerase promoter from the T7 phage, which allows you to do orthogonal regulation by expressing the T7 polymerase elsewhere in the genome. Another class changed ribosomal binding sites into the canonical sequence in order to tune up the translation initiation rates. Still another did the inverse by tuning down the translation initiation rates.

Using this pooled microarray-based approach, (MO)-MAGE is incredibly cost-effective and makes it possible to do large-scale genome engineering that would be impractical any other way. I calculated that if we had to make the equivalent number of reactions on a microarray chip using column-based synthesis, it would take something like $7 million and at least 5 months of work just to get the raw DNA. We can now do the equivalent work for a couple of thousand dollars in just a couple of weeks.

How do you know which oligos to insert in order to engineer the genome in specific ways?

If you want to target thousands of sites simultaneously, you're not going to be able to select the oligos that target them by hand. And so as we were developing (MO)-MAGE, we also worked in collaboration with Morten Sommer at the Technical University of Denmark (DTU) to develop a computational design tool called MODEST. This web-based tool allows you to upload the positions and identities of a set of mutations across the genome that you want to make, and then uses an algorithm to generate a table of all of the oligo sequences you need to produce those mutations. You can then take that table to your favorite microarray synthesis company and after a couple of weeks they will send you back a tube containing those oligos.

How does being able to generate such large numbers of targeted mutations change the research that's now possible?

The biggest problem with random mutagenesis is that the likelihood of a finding a beneficial mutation is astronomically low. (MO)-MAGE is not random, but it's not a completely rational approach to engineering either. I like to think of it as a semi-rational approach whose beauty is that by allowing you to make many genetic variants very quickly, it opens up experimental opportunities that we've never really had before.

For example, computational analysis or the scientific literature might lead you to hypothesize that 5 genes are relevant in a specific biochemical process you are trying to optimize. But those genes exist within a complex molecular system and so identifying the ideal levels for all of these components in combination using traditional approaches poses a very difficult problem. By using (MO)-MAGE, however, you can quickly produce lots of genetic variants that you can just experimentally isolate and characterize. This allows you to tune the expression of all of the genes in an iterative way.

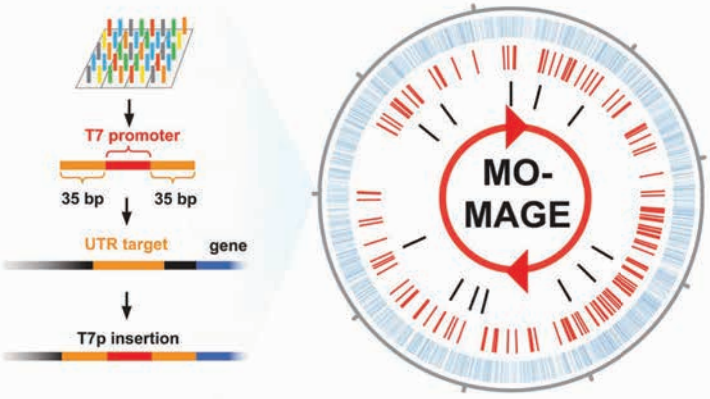
If you think about the traditional engineering pipeline that goes from design to building to testing, using this kind of semi-rational approach removes a historical bottleneck. Previously you might have been able to propose a variety of possible designs to optimize a specific biochemical activity, but it was never practical to build them all. (MO)-MAGE saves you from needing to put all your eggs in one basket with one design; it gives you a method to experimentally try hundreds of thousands or even millions of mutations and see what looks interesting. We've fixed that part of the pipeline.

In doing so, however, we've identified a new bottleneck, which is the question of how you systematically analyze the results of all of these mutations. We can build libraries of DNA very efficiently, but how do we study them in an equally efficient way? Are there genetic tricks or selection tricks that you could use to quickly pull out the most relevant mutations? These are the big questions the field is facing right now, and finding good answers would be very helpful for identifying or producing small molecules or materials of interest.

In what kinds of biological research is (MO)-MAGE going to be most useful?

What's great about using this kind of semi-random approach is that it can often lead to new, unexpected turns in biology. If you're only making one mutation at a time, you can only test one hypothesis at a time. But if you can generate 10 mutations, maybe 1 out of the 10 will give you an unexpected result. This offers a unique opportunity for further investigation.

Also, in addition to giving you information about the positive mutation space, large-scale perturbation using (MO)-MAGE also reveals the negative space. That is to say, some mutations might be really important to a particular biological problem, but because you can search the mutation space in a very comprehensive manner, you also eliminate everything that's not important. For example, if you do a saturated mutagenesis of a region of 6 base pairs — in which you permute through every single base pair variation — you will have data on more than 4000 different variants, and you've completely covered that sequence. At that point, you should be able to say that you've experimentally validated



Wang and his collaborators demonstrated the feasibility of large-scale mutagenesis by inserting T7 promoters upstream of 2585 operons in E. coli using (MO)-MAGE. For each locus, the oligonucleotide promoter sequence was incorporated in a highly targeted way between two flanking regions that then incorporated it into single-stranded DNA.

your biological model or identify ways to make the model better. (MO)-MAGE is also a particularly useful technique for understanding basic protein function because it allows you to target every single proton codon sequence one at a time. Traditionally, people have done this using alanine scans, a time consuming process in which you change every single amino acid to an alanine one at a time. If the change causes the protein to lose function you know that position is critical to the function of the protein. But using (MO)-MAGE we can now introduce pools of oligos such that within each pool you are not only scanning for alanine, but also for the other 19 amino acids. You can create a huge library of protein variants in which you assign mutations to each position, or multiple positions simultaneously, in a way that allows you to think about protein function holistically. To investigate the temperature stability of a protein, for example, you could generate a complete set of possible variants, subject them to heat stress, and then see which of those proteins remain stable.

In a project that's in the pipeline right now we are using (MO)-MAGE to target essential genes. These genes are hard to manipulate with almost any other method; you can't insert an antibiotic cassette in the middle of the essential gene because it will kill the cell. But by using these oligo pools you can essentially tile all the mutations you would be interested in making across the entire essential gene. And because each of the mutations could potentially affect the fitness of the cell, once you've made mutations across all of these positions you can then compete the variants together in a pool. The ones that are least disrupted will grow the fastest and the ones that are most disrupted will grow the slowest. You can then quantify, using deep sequencing, the relative frequencies of all of those variants across a pool. This would let you reconstruct a map of the mutation effect of this protein in one fell swoop.

Industry has taken a lot of interest in synthetic biology. What do you see as some of (MO)-MAGE's potential commercial applications?

I recently presented (MO)-MAGE at a conference and people were very excited about the prospects of doing large-scale genomic perturbations at the industrial level. It took DuPont 12 years and something like $300 million to engineer a strain of 1,3-Propanediol that's used today as an additive in many applications, from carpeting to paint thinners. They make millions of tons of this stuff in huge fermentation reactions, and to optimize the process they introduced something like 37 mutations. In contrast, we can now generate hundreds of thousands of targeted mutations for a small fraction of this time and cost.

If you're an engineer working in industry, once you've identified a biochemical pathway that produces a desired product, the next challenge is to optimize it. Traditionally, this process has been very clunky, requiring a lot of labor-intensive trial and error. Now you can just make a list of the genes and mutations that might be important and try all of the possible combinations. Each gene in one of these multi-gene pathways is like a knob, and (MO)-MAGE allows you to dial the strength of those individual knobs to identify the configuration of genetic mutations that generates the maximum flow of resources through that pathway.

Scientists and engineers who work in industry are also concerned about things like pH resistance, solvent resistance, temperature resistance, and other things that constitute global changes whose origins are unclear. Some mutations that are seen in the development of these kinds of resistance are hitchhiker mutations, while others are key driver mutations. Using (MO)-MAGE you could conceivably look at parallel strains with independent mutations and then generate a hybrid version of those two strains in order to incorporate combinations of mutations. Evolution working on its own may not have had time to access those mutations in concert, but as synthetic biologists we can. It gives us the opportunity to leapfrog to other beneficial genetic states.

What's next for (MO)-MAGE?

We are definitely interested in further developing (MO)-MAGE so it could be applied to other scientifically relevant organisms. We designed MAGE to engineer *E. coli*, but it could also potentially be used to look at the molecular origins of virulence in pathogenic *E. coli* and things of that sort.

My lab is also thinking about how these approaches could be used to engineer the gut microbiome. For example, we're thinking about strategies for generating oligos inside the cell instead of having to deliver them. If you could have the cells produce them naturally you could potentially improve the incorporation efficiency. These kinds of advances would add to the (MO)-MAGE repertoire from a technology perspective.

There's also lots of really interesting biology that could come out of this. For example, scientists have identified many mutations through experimental laboratory evolution. But the question is, are all of those mutations important? What are those mutations doing? Do those mutations have to work in concert or do they work individually? Are mutations additive or do they have synergistic or antagonistic effects? Obviously you can't "rewind" evolution, but in a certain way (MO)-MAGE offers that opportunity. We can now introduce specific mutations into an ancestral strain in any combination we want, in any order we want, creating a very directed evolution. Essentially, this gives you a way to do retro-evolution through forward engineering, letting you unwind the evolutionary process. It's an application that is ripe for this type of technology.

Related publications:

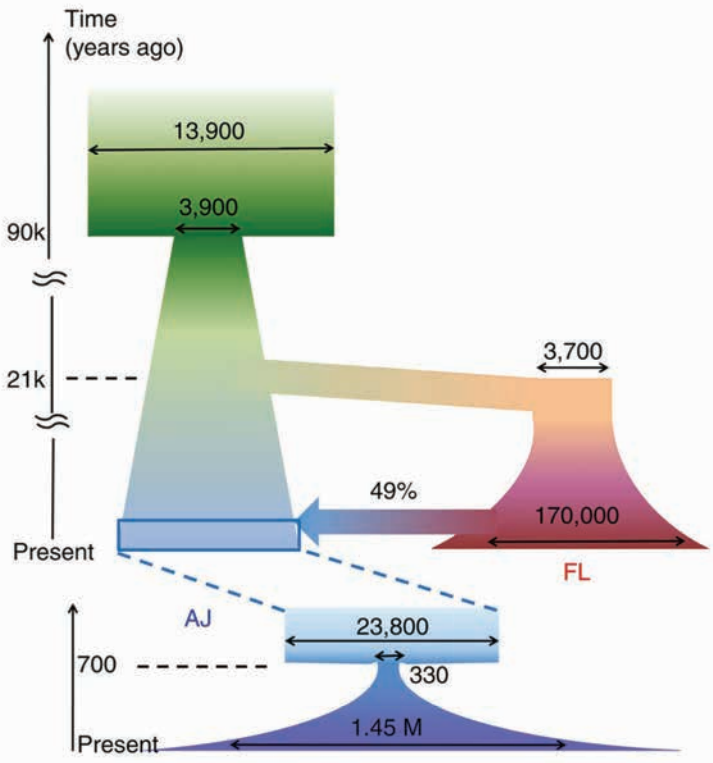
Bonde MT, Kosuri S, Genee HJ, et al. **Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides**. ACS Synth Biol. 2014 Jun 20. [Epub ahead of print]

Bonde MT, Klausen MS, Anderson MV, et al. **MODEST: a web-based design tool for oligonucleotide-mediated genome engineering and recombineering**. Nucleic Acids Res. 2014 Jul;42:W408-15.

Study Sheds Light on Ashkenazi Jewish Genome and Ancestry

An international research consortium led by Associate Professor Itsik Pe'er has produced a new panel of reference genomes that will significantly improve the study of genetic variation in Ashkenazi Jews. Using deep sequencing to analyze the genomes of 128 healthy individuals of Ashkenazi Jewish origin, The Ashkenazi Genome Consortium (TAGC) has published a resource that will be much more effective than previously available European reference genomes for identifying disease-causing mutations within this historically isolated population. Their study also provides novel insights into the historical origins and ancestry of the Ashkenazi community. A paper describing their study was published online in *Nature Communications*.

The dataset produced by the consortium provides a high-resolution baseline genomic profile of the Ashkenazi Jewish population, which they revealed to be significantly different from that found in non-Jewish Europeans. In the past, clinicians' only option for identifying disease-causing mutations in Ashkenazi individuals was to compare their genomes to more heterogeneous European reference sets. This new resource accounts for the historical isolation of this population, and so will make genetic screening much more accurate in identifying disease-causing mutations.



The consortium's model of Ashkenazi Jewish ancestry suggests that the population's history was shaped by three critical bottleneck events. The ancestors of both populations underwent a bottleneck sometime between 85,000 and 91,000 years ago, which was likely coincident with an Out-of-Africa event. The founding European population underwent a bottleneck at approximately 21,000 years ago, beginning a period of interbreeding between individuals of European and Middle Eastern ancestry. A severe bottleneck occurred in the Middle Ages, reducing the population to under 350 individuals. The modern-day Ashkenazi community emerged from this group.

In an article that appears on the website of Columbia University's Fu Foundation School of Engineering and Computer Science, Dr. Pe'er explains, "Our study is the first full DNA sequence dataset available for Ashkenazi Jewish genomes... With this comprehensive catalog of mutations present in the Ashkenazi Jewish population, we will be able to more effectively map disease genes onto the genome and thus gain a better understanding of common disorders. We see this study serving as a vehicle for personalized medicine and a model for researchers working with other populations."

In addition to offering an important resource for such future translational and clinical research, the paper's findings also provide new insights that have implications for the much debated question of how European and Ashkenazi Jewish populations emerged historically.

By analyzing the proliferation of long nucleotide sequences that are identical in the unrelated individuals in their sample, the researchers determined that a population bottleneck of approximately just 350 Ashkenazi Jewish individuals occurred in central Europe about 700 years ago, followed by an exponentially rapid population increase. The findings suggest that the ancestry of all present-day Ashkenazi Jews can be traced back to this small population.

Moreover, by comparing the genomes of Ashkenazi Jews with those of Flemish origin, the researchers found strong evidence that the ancestry of the modern-day Ashkenazi can be traced to a fairly even mixture of European and Middle Eastern descent. Cross-breeding appears to have occurred at approximately the same time as the Ashkenazi bottleneck, suggesting that when Jewish migrants arrived in Europe from the Levant they mixed with the local population.

The consortium's findings also indicate that modern European ancestry can be traced to migration from the Middle East following the last great ice age, approximately 12,000-25,000 years ago. Others have argued that Europe was populated by an earlier migration from Asia more than 40,000 years ago, though this study provides strong evidence against that hypothesis.

The whole genome sequence data produced by this study are available at the European Genome-Phenome Archive. Dr. Pe'er anticipates that it will become a resource for other scientists interested in population genetics and personalized medicine.

Related publication:
Carmi S, Hui KY, Kochav E, et al. **Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins**. Nat Commun. 2014 Sep 9;5:4835.

Distinguishing Patterns of Tumor Evolution in Chronic Lymphocytic Leukemia

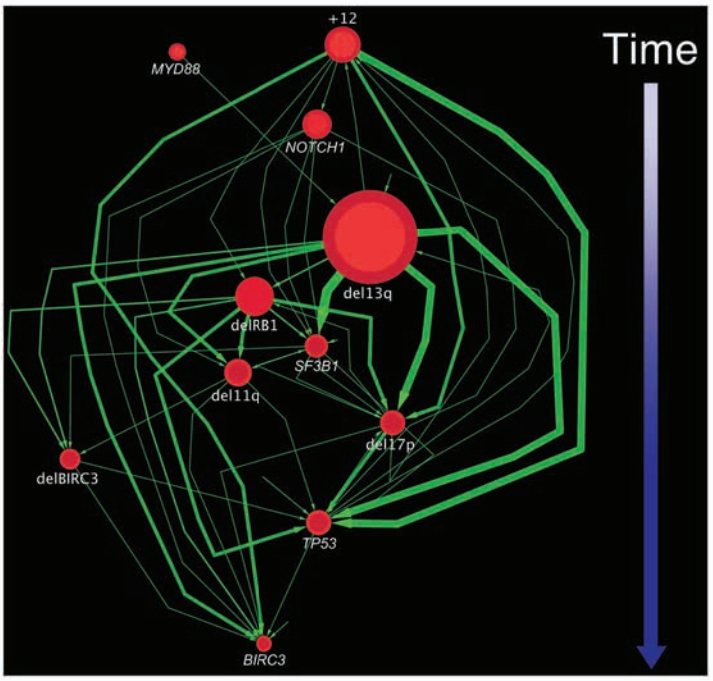
As biologists have gained a better understanding of cancer, it has become clear that tumors are often driven not by a single mutation, but by a series of genetic changes that correspond to particular stages of cancer progression. In this sense, a tumor is constantly evolving, with different groups of cells that harbor distinctive mutations multiplying at different rates, depending on their fitness for particular disease states. As the search for more effective cancer diagnostics and therapies continues, one key question is how to disentangle the order in which mutations occur in order to understand how tumors change over time.

In a paper published in the journal *eLife*, a team of investigators led by Associate Professor Raul Rabadan reports on a new computational strategy for addressing this challenge. Their framework, called tumor evolutionary directed graphs (TEDG), considers 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 (CLL), they were able to develop a model of how the disease's mutational landscape changes from its initial onset to its late stages.

The design of the approach was led by postdoctoral research scientist Jiguang Wang and associate research scientist Hossein Khiabanian, both members of the Rabadan Lab. Their method integrates a "longitudinal" analysis — looking at changes in the mutational landscape of a single tumor across multiple time points — as well as a "cross-sectional" analysis using statistical models based on tumors collected from large numbers of patients. For each tumor, if one genetic change is observed before another, a graph connects them with an edge that represents their sequential order. The investigators then consider many sequential networks from different patients to create what they call an Integrated Sequential Network (ISN). The final TEDG can then be inferred from the ISN by removing indirect associations using a technology called network deconvolution (recently developed at the Massachusetts Institute of Technology). The final graph shows the relationships among the different cancer-driving gene mutations, with arrows indicating the order in which they typically arise.

The authors chose CLL to test their approach because it progresses relatively slowly, and blood samples containing leukemic cells can be collected regularly without causing patients any major discomfort. The study used targeted deep sequencing and fluorescence in situ hybridization (FISH) data gathered from 70 CLL patients over a period of 12 years.

When the analysis was complete, 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 authors report that the mutation sequences indicated in the resulting evolutionary graph are consistent with earlier findings, including previous work in the Rabadan Lab that determined that mutations in the gene *TP53* are relatively rare early in the course of disease, but multiply in great numbers in patients who develop chemotherapy-resistant CLL.



A graph representing the sequence of genomic alterations in chronic lymphocytic leukemia (CLL).

Dr. Rabadan believes 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. The case of *TP53*, for example, suggests that rare subpopulations of cells that might be missed in other types of studies could actually constitute better biomarkers for predicting the course an individual tumor might take. Rabadan explains, "The presence of a *TP53* mutation indicates a bad prognosis, but by the time it becomes visible in sufficient amounts, it is usually too late to address it effectively. If at the initial time of diagnosis we could accurately predict the evolutionary trajectory that leads to this kind of activity, we might be able to identify other biomarkers that arise earlier and predict that the disease will be resistant to chemotherapy. This would tell you that you need to treat the cancer more aggressively." Although the findings in this study are just a beginning in the effort to achieve this kind of application in the clinic, Rabadan points out that they open the door to a new way of analyzing how complex diseases evolve.

Related publication:

Wang J, Khiabanian H, Rossi D, et al. **Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia**. Elife. 2014 Dec 11;3.

Systems Biology Throws You Out of the Box: A Conversation with Saeed Tavazoie

One of the defining features of systems biology has been its integration of computational and experimental methods for probing networks of molecular interactions. The research of Saeed Tavazoie, a professor in the Columbia University Department of Systems Biology, has been emblematic of this approach. After undergraduate studies in physics, he became fascinated by the processes that govern gene expression, particularly in understanding how gene expression is regulated by information encoded in the genome. Since then, his multidisciplinary approach to research has generated important insights into the principles that orchestrate genome regulation, as well as a number of novel algorithms and technologies for exploring this complex landscape.

In this conversation, Dr. Tavazoie discusses his research in the areas of gene transcription, post-transcriptional regulation, and molecular evolution, as well as some innovative technologies and experimental methods his lab has developed.

How would you describe your approach to biological research?

Traditionally in biology, you had to take a very focused view of a problem. You might look closely at a particular protein or a specific biological process, and after reading a lot of papers about it you could begin asking a question and get an answer that moved our understanding of that problem a little bit forwards. In systems biology, however, we are not interested in one particular protein or process, but in identifying the underlying principles of how the entire system behaves. This perspective has been made possible by a number of new technologies that now allow us to make large-scale measurements on the entire system — including not just all of the genes and proteins, but all of the interactions between them. Once we gather these large collections of data, we can then analyze them to reverse engineer how all of the components within a regulatory network come together to orchestrate cellular behavior.

This approach brings about an important change in perspective. Instead of imposing a set of expectations that define how we think biology should behave, we try to be as unbiased as possible and let the system tell us what's interesting. This isn't to say that the traditional methods aren't important, but when an entire system behaves in ways you never anticipated, that's when important new discoveries are made. Although it can be useful to explain how one molecule interacts with one other molecule, the goal should be to walk away from your experiment learning a general principle that goes well beyond this specific interaction. In almost every experiment we do in our lab we try to set things up so that it could potentially produce these kinds of insights.

The new technologies that we and others are developing not only generate a scaffold of knowledge about regulatory interactions that other scientists can use, but eventually become important tools for making progress in many other areas. In the past, technologies like microarrays, RNA-Seq, and CHiP-Seq totally changed the way people do science. Today, new technologies that are coming out of systems biology are pushing conceptual revolutions in biology because they enable you to make observations you couldn't make before. It's not just that you start thinking out of the box, new technologies actually throw you out of the box and you can't avoid thinking about things in new ways.

Can you give an example of a technology that you are working on and how it is changing your perspective on biology?

One area of research that has seen a lot of activity in recent years is to study how transcription factors regulate gene expression. Transcription factors are proteins that bind to DNA and either promote or repress DNA's transcription into RNA. There is a technology called CHiP-Seq that uses antibodies that can recognize and bind to specific transcription factors. You can fragment the DNA into millions of pieces and then use an antibody to find all the places in the genome that are bound by this protein, giving you a snapshot of whether or not a particular gene is being regulated at a particular time point.

Although this is a powerful approach, it can only tell you the binding locations of a single kind of transcription factor at a time. Some years back I wondered if it might be possible to develop a technology that would let us identify the binding sites for all of the hundreds of proteins that are bound to different segments of the genome at a given time, all at once. In our lab we have now developed a technology that can do this. Using a series of purifications of protein-DNA complexes, we can separate naked DNA from DNA that has proteins bound to it, process these segments, and then use high-throughput sequencing to identify the chromosomal location for each of the bound transcription factors, giving a global profile of protein occupancy throughput the genome. This technology, co-developed with a former graduate student, Tiffany Vora, allows us to identify all the sites in the genome that are bound by protein, but by itself can't tell you the identity of each protein. To solve this problem, Peter Freddolino, a postdoctoral re-

searcher in my lab, developed a computational method that takes this readout of protein occupancy and, using a modification of an algorithm called FIRE that my lab developed previously, generates binding sequence models of where the proteins are found. So now you can go in and systematically learn from these experiments what the sequence preferences are for a large number of proteins at once.

Having this technology opens up a number of exciting possibilities, particularly when it is combined with RNA-Seq, which gives a complete picture of the RNAs in the cell at a given time. If you perturb the cell in some way it now becomes possible to monitor all of the locations in DNA that are bound by transcription factors, while simultaneously recording the entire profile of RNAs that are present (also called the transcriptome). Because the networks of interactions among these different components are dynamic, we take a series of measurements every few minutes, monitoring how these interactions evolve over time. We want to get to the point where you could basically make a movie consisting of a series of snapshots of which proteins are binding to which genes and when the RNAs are being made. It would be really enabling for systems biology, giving us the kind of observations that we really want. And most importantly, it would be mechanistically anchored in physical interactions.

What would be some of the potential applications of this kind of approach?

In our initial tests, we focused on *E. coli*, a bacterium that has been studied for about 70 years and whose binding specificities are largely known. When we use our approach without using any of these earlier findings as input, we rediscover those binding sites that people discovered over decades. But whereas it required years of intensive laboratory work to find these binding events in the past, we identify them, and others, in just a few experiments.

Now that we know this works in *E. coli*, we can begin learning binding site preferences in other organisms we don't know anything about. The vast majority of bacteria that cause infectious diseases and play important roles in ecosystems have not been studied at the level of detail we now have for *E. coli*, so we know very little about their regulatory networks. We can now take bacteria that are important but not well studied, run them through our pipeline, systematically annotate all of their binding sites, and generate binding site models that would rapidly expand our knowledge about their regulatory networks.

That kind of result would be significant in and of itself, but our approach also produces knowledge about regulatory interactions inside cells that will be useful to scientists who work in pathogenic systems, no matter what they're studying. These regulatory networks are engaged in almost any process someone might look at, and react to any kind of perturbation you might study, like oxidative stress or antibiotic stress. Knowing the mechanisms through which regulatory networks function is a huge step toward figuring out what's going on. This will be really powerful.

How about other layers of regulation, such as post-transcriptional regulation?

Although a big focus for systems biology in its first 10-15 years has been on transcription, we and others are discovering that there's also a huge amount of regulation that occurs after the messenger RNA (mRNA) has been created. In the regions at the ends of the transcript called 5' and 3' untranslated regions, for example, the mRNA can be bound by proteins that can increase the degradation rate of the mRNA. Discovering these elements in RNA is more challenging than in DNA because RNA is single stranded and can form secondary structures where the proteins bind. All of this means that even if studies along the lines of what I was talking about earlier determine that you are highly expressing a gene at a particular moment, the transcript could be degraded very quickly and might not actually play the role that those findings might indicate.

Over the last few years we have developed computational methods that look through the entire transcriptome and discover regulatory elements in RNA. Basically, we simulate RNA binding by testing hundreds of billions of possible structures in the computer and calculate which ones have a high likelihood of being involved in post-transcriptional regulation. We need massive computational resources to sift through all of them and identify ones that are functionally important. To do one run on a data set can take 3-4 days using 400 processors on the Department of Systems Biology's high-performance computing cluster. That's a huge amount of computing that would be hard to do without having this kind of infrastructure here at Columbia.

We're finding that post-transcriptional regulation plays important roles almost everywhere we look. The challenge now is not only to catalog the protein binding sites, but also to find the proteins that recognize them, figure out what the proteins are doing, and see what other regulators a protein is interacting with; this knowledge is necessary to work out the entire pathway of regulation.

It's exciting because if you think about cancer biology, for example, over the years people have discovered that a large number of oncogenes and tumor suppressors are transcription factors. But we're discovering that gene expression is modulated not only by how much RNA you make, but also by how fast the mRNAs degrade, because if you degrade an mRNA it doesn't have a chance to generate a protein. There are actually two regulatory inputs, and understanding how they interact is going to provide a clearer picture of what's happening at the molecular level. Much of post-transcriptional regulation is still a black box, though, because people have not been able to study it effectively so far. We're trying to change that.

Do these regulatory networks tend to be stable, or do they change over time?

Our lab is actually very interested in the principles that govern how regulatory networks change over very long evolutionary time scales. We study this by carrying out experimental evolution in the laboratory. The nice thing about working with bacteria is that they divide once every hour, so over the course of weeks to months, you see a large number of generations. In our experiments we can expose them to different extreme environments and see how cells survive and adapt to the challenges that the new environments create, look at what kinds of modifications occurred in the ge-

nome, and determine how they affected the regulatory network. One of the things we've discovered is that it seems to be much easier than we had previously thought for bacteria to adapt to extreme environments, and they seem to do so by rewiring their regulatory network. Historically, most people have thought that adaptation to new environments happens by making changes in the coding regions of the genome and that these changes generate a protein that functions better in the new environment, improving the organism's fitness. People have thought that these kinds of positive improvements in fitness and adaptation happen through very gradual, subtle, rare changes in amino acids at the protein level and take a long time to happen.

What we're discovering, though, is that upon the transition to these extreme environments, the dominant mutation mechanisms are not the gain of new functions through advantageous mutations, but rather loss-of-function mutations in genes that play regulatory roles. That is, the organisms adapt when one gene is lost that lets another gene that it had previously suppressed become activated. That's the nature of regulatory networks.

Although it wasn't initially obvious, this makes sense because it's very easy to mutate a gene in ways that will make a protein stop functioning. It's much, much harder to get subtle, rare mutations that actually enable the bacteria to survive better. This is because the rich regulatory network that exerts control over all the genes evolved in the native habitat of the organism. When you take the organism outside that context, there's no reason to expect that the regulatory network is going to do the right thing. It's like taking an organism with a cognitive system and putting it in a completely weird psychological environment. It's going to go crazy. It's not going to adapt. That's the way we think about it now. And we're seeing a huge amount of this, which is very exciting.

We're exploring how this perspective could help to explain antibiotic resistance. Within the clinical setting, there has been great concern because, increasingly, bacteria can survive treatment with antibiotics that were once able to eliminate infections. We would like to identify the major antibiotic mechanisms by discovering how bacteria adapt to environmental challenges and evolve novel pathways of resistance.

You compared this exposure to extreme environments as producing a cognitive response. That's not to say that bacteria can think, is it?

Obviously they can't think in the same way that humans think, but a few years back we applied a systems biology approach to looking at bacterial behavior. We discovered that bacteria make predictions about their environment, very much like nervous systems do. They actually anticipate what's going to happen next.

When we raised the temperature at a constant concentration of oxygen on a culture of *E. coli*, we saw that all the genes involved in aerobic metabolism went down in a synchronized way. We initially found this nonsensical because it isn't advantageous to shut down aerobic metabolism in the presence of oxygen. But the changes make sense in the context of the bacteria's native habitat if you look at the ecology of the organism and what it's been experiencing over geological time scales. When *E. coli* enter the mouth, the temperature rises to 37°C. Then they go down the gastrointestinal tract and after 10-20 minutes oxygen levels drop. So when the temperature goes up, the networks start working in anticipation of the change in oxygen levels. This kind of anticipatory behavior was not known in cellular systems before.

The only way we could see this was by looking at the entire gene expression dataset, and so this is a great example of the power of the systems biology approach. Because it's minimally biased it puts you in a domain of exploration that you couldn't have anticipated ahead of time. It's not unlike the bacteria, actually, in that scientifically you get dropped into an environment you haven't seen before and need to adapt your perspective to what the data are telling you. This approach has changed how we do biology in ways that are very exciting to be a part of.

Kyle Allison Receives NIH Early Independence Award

Kyle Allison has received the NIH Director's Early Independence Award. This program enables outstanding young investigators who have recently completed their PhD's to move rapidly into independent research positions. In combination with the Department of Systems Biology Fellows program, this grant has allowed him to open his own laboratory at Columbia and pursue independent research to investigate the problem of bacterial persistence.

Bacterial persistence occurs when individual bacterial cells survive treatment with antibiotics that kill other cells that are genetically identical, and do not have a mutation. It is a particular problem in human health, as chronic, drug-resistant bacterial infections affect thousands of Americans and cost millions of dollars to treat annually. Scientists do not fully understand the causes of bacterial persistence, though its occurrence suggests that bacterial populations are not homogeneous but contain subpopulations of cells that are physiologically distinct.

With the support of his Early Independence Award, Dr. Allison's goal is to develop methods that would make it possible for the first time to study such bacterial heterogeneity at the level of single cells. Using a combination of experimental and computational approaches based in systems biology, he will develop methods for isolating persistent cells and characterizing them at the molecular level. These techniques should provide deep insight into what distinguishes persistent cells from close relatives that are sensitive to antibiotics, and could accelerate development of new antibiotic treatments.

Synergy between Two Genes Drives Aggressive Prostate Cancer

Two genes work together to drive the most lethal forms of prostate cancer, according to research by investigators in the Columbia University Department of Systems Biology. These findings could lead to a diagnostic test for identifying those tumors likely to become aggressive and to the development of novel combination therapy for the disease.

The two genes—*FOXM1* and *CENPF*—had been previously implicated in cancer, but none of the prior studies suggested that they might work synergistically to cause the most aggressive form of prostate cancer. "Individually, neither gene is significant in terms of its contribution to prostate cancer," said co-senior author Andrea Califano, Chair of the Department of Systems Biology. "But when both genes are turned on, they work together synergistically to activate pathways associated with the most aggressive form of the disease."

"Ultimately, we expect this finding to allow doctors to identify patients with the most aggressive prostate cancer so that they can get the most effective treatments," said co-senior author Cory Abate-Shen, also a member of the Department of Systems Biology. "Having biomarkers that predict which patients will respond to specific drugs will hopefully provide a more personalized way to treat cancer."

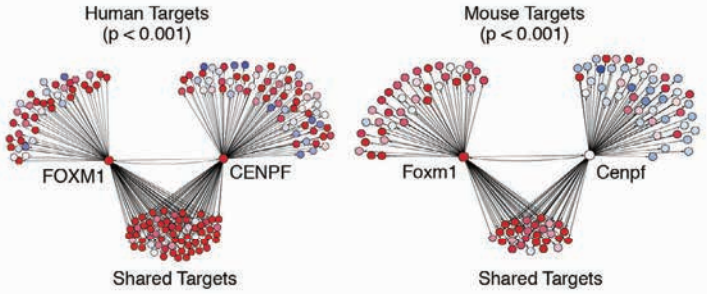
Scientists widely recognize that cancer is characterized by multiple genetic changes. "However, distinguishing the handful of genes that are driving the cancer from the many genes whose altered expression does not contribute directly to the cancer has proven to be a daunting task," said Dr. Califano. "It becomes even more difficult when genes work together synergistically, because they must be analyzed in pairs rather than one by one. This results in an enormous number of possible combinations that defies our best statistical tools and requires sophisticated systems biology approaches."

"Prostate cancer is particularly challenging because it has such a wide variety of clinical presentations, with relatively few shared genetic mutations," said Dr. Abate-Shen. Thus, to find the key genes that drive prostate cancer, the CUMC team devised a novel experimental approach in which they used computational approaches to compare the gene regulatory networks that drive prostate cancer in humans with those in a genetically engineered mouse model of the disease.

In many cancer studies, researchers rely on mouse models to identify genes that are expressed in disease. "However, inherent differences between species often make it difficult to extrapolate findings in mice to humans," said Dr. Abate-Shen.

"By tracing the regulatory logic of these tumors in both species," said Dr. Califano, "we were able to identify identical driver genes of malignant prostate cancer and to discover that they don't work as individual drivers but rather together, as a synergistic driver pair."

Using the high-performance computing cluster housed in the Department of Systems Biology, the analysis determined that *FOXM1*



Computational synergy analysis depicting FOXM1 and CENPF regulons from the human (left) and mouse (right) interactomes showing shared and nonshared targets. Red corresponds to overexpressed targets and blue to underexpressed targets.

and *CENPF* jointly control genetic programs associated with the most prominent tumor hallmarks in both species. Individually, the aberrant expression of these genes does not activate these programs. When acting together, however, the two genes can wreak havoc in the cancer cell and turn it into a very aggressive tumor.

To validate the roles of *FOXM1* and *CENPF*, the researchers silenced the expression of the genes in four human prostate cancer cell lines, first individually and then together. Silencing either gene alone had only a modest effect on the ability of the cells to form tumors. However, co-silencing both genes at once completely stopped the growth of tumors in a mouse. This observation is consistent with a synergistic interaction, where the joint effect of both genes is much greater than the sum of their individual effects.

The researchers then analyzed prostate cancers from a group of more than 900 patients who had undergone prostate removal surgery. This analysis showed a striking correlation between the co-expression of *FOXM1* and *CENPF* and the poorest disease outcome. In sharp contrast, expression of either gene alone did not correlate with aggressive disease. In addition, tumors in which neither gene was aberrantly expressed had the best prognosis. The researchers also showed that silencing the two genes inactivated pathways known to be hallmarks of aggressive prostate cancers, suggesting the possibility that combined therapeutic targeting of both *FOXM1* and *CENPF* could arrest human disease," said Dr. Abate-Shen.

"This is just a first step toward a deeper understanding of the genetics of cancer," said Michael Shen, also a member of the Department of Systems Biology and a contributor to the study. "These tools and approaches may have broad utility in studying prostate cancer; cross-species computational analyses also could be used to identify the causes of other cancers, as well as that of other complex diseases."

Related Publication:

Aytes A, Mitrofanova A, Lefebvre C, 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.

Grants, Awards, and Honors

Dimitris Anastassiou was elected to the National Academy of Inventors, a distinction awarded to academic inventors who have created or facilitated outstanding inventions that have made a tangible impact on quality of life, economic development, and the welfare of society.

Andrea Califano was named a recipient of the HICCC Inter-Programmatic Award, for a collaboration with Matthew Maurer focusing on identifying master regulator-directed therapy for patients with residual breast cancer following neoadjuvant chemotherapy. The Califano Lab also received a grant from SWOG to support N-of-1 clinical trials. These grants have made it possible to open a new breast cancer combination trial that has begun enrolling patients.

Oliver Hobert was named a AAAS Fellow. This honor recognizes AAAS members who have made significant contributions to science, based on recognition from their peers.

Nathan Johns, a graduate student in Harris Wang's lab and the Integrated Program in Cellular, Molecular and Biomedical Studies, was awarded the National Science Foundation Research Fellowship. The program recognizes and supports outstanding graduate students in NSF-supported science, technology, engineering, and mathematics disciplines who are pursuing research-based master's and doctoral degrees at accredited US institutions.

Dana Pe'er is a recipient of one of five inaugural Stand Up to Cancer (SU2C) Phillip A. Sharp Innovation in Collaboration Awards. The prizes are designed to encourage current SU2C scientists to explore synergistic and innovative collaborations that will further enhance the mission to accelerate new cancer treatments.

Raul Rabadan was awarded tenure in the Department of Biomedical Informatics and Department of Systems Biology. He was promoted to Associate Professor.

Michael Shen is a co-PI of a multi-institution Prostate Cancer Foundation (PCF) Global Treatment Sciences (GTSN) Challenge Award that will investigate whether prostate tumors with *SPOP* mutations are selectively susceptible to DNA-damaging therapeutic agents.

Brent Stockwell was named a recipient of the 2014 Lenfest Distinguished Columbia Faculty Award, which recognizes excellence in the teaching and mentoring of undergraduate and graduate students.

Nicholas Tatonetti received a PhRMA Foundation Early Career Award, which provides support to individuals beginning independent research careers in academia.

Harris Wang received an NSF CAREER award. The award will support research to use systems and synthetic biology methods to experimentally probe the rules determining horizontal gene transfer between diverse bacterial species.

Dennis Vitkup received a multiyear grant from the National Institute of General Medical Sciences (NIGMS) to develop models of metabolic networks for all of the major bacterial species that cause disease in humans.

Chaolin Zhang received an Explorer Award from the Simons Foundation Autism Research Initiative. His project is focused on modeling alteration of the RBFOX1 (A2BP1) target network in autism.

PhD Graduates

Congratulations to all Department of Systems Biology students who successfully defended their PhD theses in 2014:

Zachary Carpenter (Rabadan and Ferrando Labs)
Wei-Yi Cheng (Anastassiou Lab)
Mariam Konaté (Vitkup Lab)
Anat Kreimer (Itsik Pe'er Lab)
Allan Lazarovici (Bussemaker Lab)
Eugenia Lyashenko (Vitkup Lab)
Felix Sanchez-Garcia (Dana Pe'er 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

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