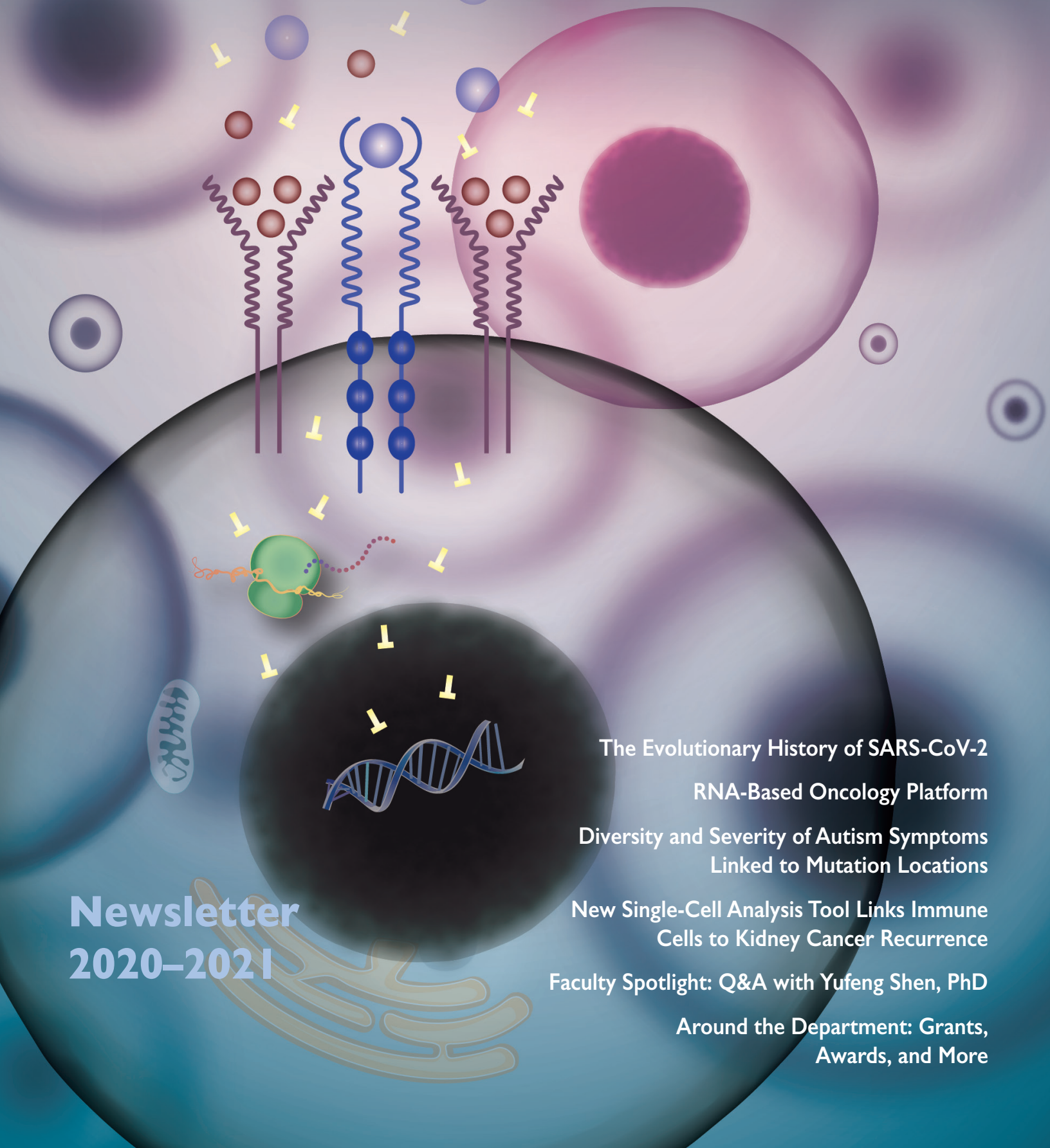
Department of Systems Biology

Columbia University Irving Medical Center



Newsletter 2020–2021

The Evolutionary History of SARS-CoV-2

RNA-Based Oncology Platform

Diversity and Severity of Autism Symptoms Linked to Mutation Locations

New Single-Cell Analysis Tool Links Immune Cells to Kidney Cancer Recurrence

Faculty Spotlight: Q&A with Yufeng Shen, PhD

Around the Department: Grants, Awards, and More

COLUMBIA UNIVERSITY
IRVING MEDICAL CENTER
Department of Systems Biology

Table of Contents

Highlights 2020–2021

Inside this issue, learn about the latest research, noteworthy grants and milestones from the Department of Systems Biology.

The Evolutionary History of SARS-CoV-2

Structural modeling indicates that ancestors of SARS-CoV-2 may have acquired the ability to infect humans decades ago. ..... **Page 3**

RNA-Based Oncology Platform

A framework based on a Master Regulator conceptualization of cancer regulation has the potential to radically expand precision cancer treatment by providing rapid prioritization of available drugs. ..... **Page 5**

Diversity and Severity of Autism Symptoms Linked to Mutation Locations

De novo truncating mutations, which are responsible for close to 5% of autism cases and up to 20% of cases seen clinically, occur in parents' germline cells and usually strongly disrupt the functions of target genes. ..... **Page 8**

New Single-Cell Analysis Tool Links Immune Cells to Kidney Cancer Recurrence

A rare and previously unknown type of immune cell in kidney tumors can predict which patients are likely to have a cancer recurrence after surgery. ..... **Page 9**

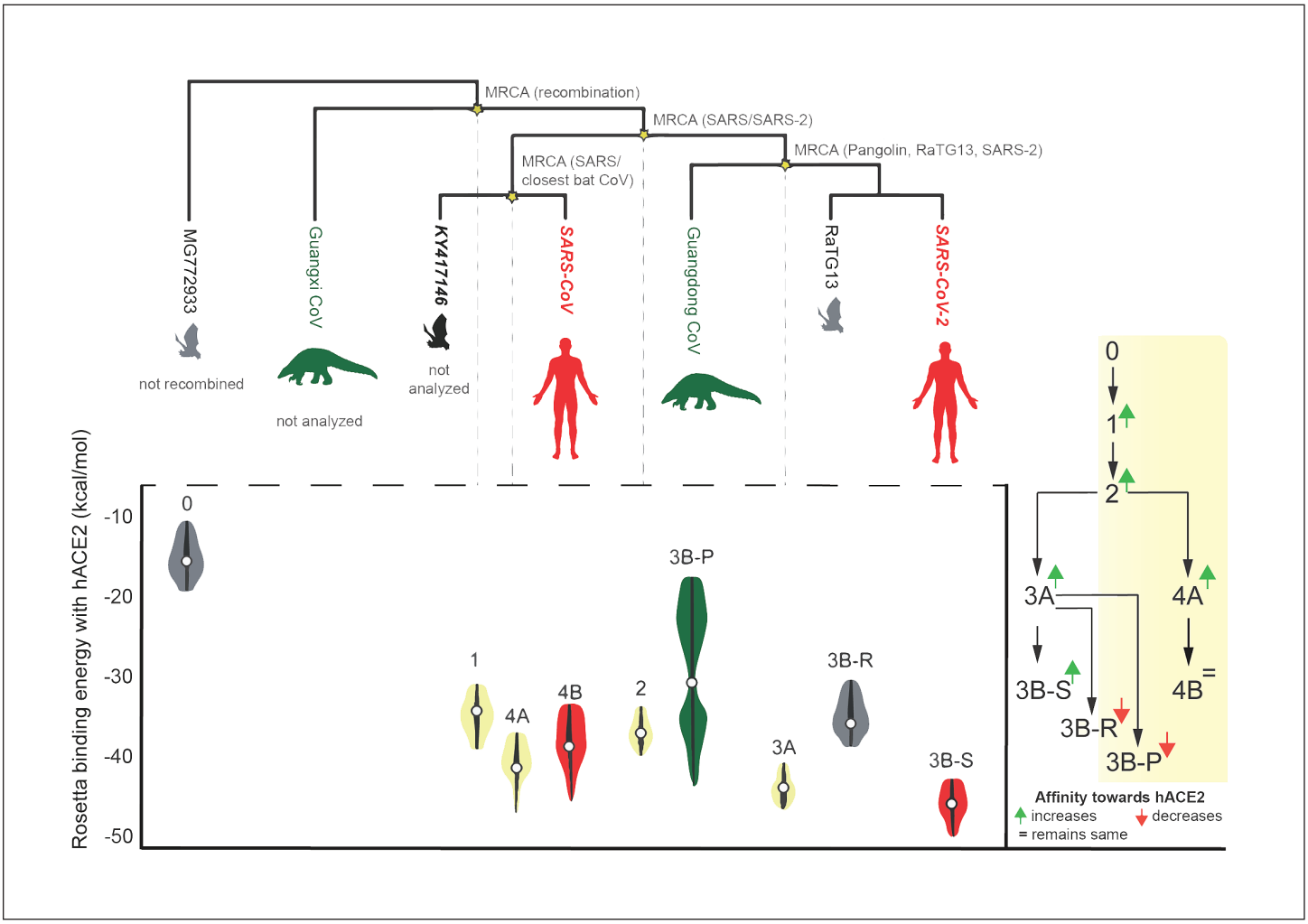
Faculty Spotlight: Q&A with Yufeng Shen, PhD

Computational biologist Yufeng Shen, PhD, studies the links between genetic variants and human diseases. ..... **Page 11**

Around the Department

Grants, Awards and More. ..... Page 12

The Evolutionary History of SARS-CoV-2



The appearance of SARS-CoV-2 underscores the need to better understand the evolutionary processes that drive the emergence and adaptation of zoonotic viruses in humans. Systems Biology members Raul Rabadan, Mohammed AlQuraishi, and Juan Ángel Patiño-Galindo address the question in a recent *Genome Medicine* paper, "Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2."

The causative agent of COVID-19, SARS-CoV-2, was a previously unknown RNA coronavirus (CoV) of the *Betacoronavirus* genus, with 80% similarity at the nucleotide level to the severe acute respiratory syndrome coronavirus (SARS-CoV), the agent responsible for the 2002–2003 SARS outbreak. SARS-CoV and SARS-CoV-2 are still the only members of the *Sarbecovirus* subgenus of *Betacoronavirus* known to infect humans. The new results suggest that recombination was a key factor in the emergence of *Sarbecoviruses* in humans.

In the *Betacoronavirus* genus, which also includes SARS-CoV and MERS-CoV, recombination frequently encompasses the receptor binding domain (RBD) of the Spike protein, which is responsible for viral binding to host cell receptors. In the current work, Rabadan, AlQuraishi, and Patiño-Galindo reconstruct the evolutionary events that accompanied the emergence of SARS-CoV-2, with a special emphasis on the RBD and its adaptation for binding to its receptor, the human ACE2 protein (hACE2), the port of entry of SARS-like viruses into human cells.

They analyze the evolution of SARS-CoV-2 and its closest relatives, with a focus on the RBD region of the Spike protein, as a means to better understand viral tropism. It was earlier hypothesized that recombination and rapid evolution has occurred in bat, civet, and human SARS-CoVs. However, previous descriptions of recombination in the Spike protein were purely observational. In contrast, Rabadan, AlQuraishi, and Patiño-Galindo use statistical methods to show that recombination events preferentially affect the Spike gene, both

at the genus level *(Betacoronavirus)* and within individual species (such as MERS-CoV).

By means of phylogenetic and recombination analyses, they found evidence of a recombination event in the RBD involving ancestral lineages of both SARS-CoV and SARS-CoV-2. They then assessed the effect of this recombination at the protein level by reconstructing the RBD of the closest ancestors to SARS-CoV-2, SARS-CoV, and other *Sarbecoviruses,* including the most recent common ancestor of the recombining clade. They used the resultant information to measure and compare, in silico, their ACE2-binding affinities using the physics-based trRosetta algorithm.

The researchers found an ancestral recombination event affecting the RBD of both SARS-CoV and SARS-CoV-2 that was associated with an increased binding affinity to hACE2. Structural modeling indicated that ancestors of SARS-CoV-2 may have acquired the ability to infect humans decades ago. The binding affinity with the human receptor would have been subsequently boosted in SARS-CoV and SARS-CoV-2 through further mutations in RBD.

To understand how recombination contributes to the evolution of *Betacoronaviruses* across different viral subgenera and hosts, they analyzed 45 *Betacoronavirus* sequences from the five major subgenera that infect mammals (*Embevovirus, Merbecovirus, Nobecovirus, Hibecovirus,* and *Sarbecovirus*). Enrichment analysis showed that recombination often involves the N-terminus of the Spike protein, which includes the RBD. Enrichment for recombination events persisted even after they restricted the analysis to the most common host (bats), suggesting that recombination was not driven simply by sampling of multiple human sequences. They concluded that recombination in *Betacoronaviruses* frequently involves the Spike protein across viral subgenera and hosts.

Thus, evolutionary analyses and structural modeling suggest that the evolutionary processes giving rise to SARS-CoV-2 included a recombination involving ancestors of SARS-CoV and SARS-CoV-2, followed by the accumulation of point mutations in the Spike protein. Both the ancestral recombination event and the point mutations, which differ between SARS-CoV and SARS-CoV-2, would have resulted in progressively tighter binding to hACE2. It appears that ancestors to SARS-CoV-2, with the ability to bind tightly to hACE2 and thus potentially infect humans, may have been circulating in the wild for decades prior to making the jump to humans and causing pandemic disease. These results show the importance of combining evolutionary analyses with protein structure and binding affinity predictions, to assess the host-switching potential of animal-infecting viruses based on the genetic changes that have accumulated along their evolution.

> "In a way, we're doing archaeology, reconstructing the events that led to the pandemic, to get information about how these events occur." —*Raul Rabadan*

While humans take 20 to 30 years to produce a new generation, viruses take a mere few hours. Viruses also lack most of the mechanisms to correct replication mistakes. As a result, the rate of error, or mutation, is higher. Given the rapid rate of viral mutation, some people have questioned why pandemics don't occur more often. The answer is a combination of conditions and chance. First of all, the virus must jump from the animal to humans. If a virus originates in bats, for example, a person in contact with the bats must become infected and then infect others. Then at least some of the infected people need to spend time in a heavily populated area such as a large city. The swine flu epidemic in Mexico City is an example of such a chain of events.

"In a way," says Rabadan about the current work, "we're doing archaeology, reconstructing the events that led to the pandemic, to get information about how these events occur."

"In the future," he adds, "more pandemics are likely to occur. Having the technology to look at the virus genomes will enable scientists to predict the risks of particular mutations, understand the mechanism of adaptation, and potentially develops means of combating the emerging threat, like vaccines."

REFERENCE

Patiño-Galindo JA, Filip I, Chowdhury R, Maranas CD, Sorger PK, AlQuraishi M, Rabadan R. Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2. *Genome Med*. 2021 Aug 6;13(1):article no. 124. doi: 10.1186/s13073-021-00943-6.

RNA-Based Oncology Platform

Andrea Califano

The ultimate objective of precision cancer medicine (PCM) is to use molecular-level properties of a tumor—such as gene expression, epigenetic modification, proteomics, and mutational profiles—to predict sensitivity to available therapeutic agents or to guide development of novel ones. Andrea Califano, Dr, chair of Systems Biology, and colleagues have developed a framework based on a Master Regulator (MR)-based conceptualization of cancer regulation that has the potential to radically expand PCM treatment by providing rapid prioritization of effective drugs from those available.

Current PCM takes primarily two complementary approaches. The first, oncogene addiction, identifies targeted therapies based on the presence of mutations that induce aberrant activity in druggable oncoproteins. The second, immunotherapy, is based on the discovery that some tumor-initiated immunosuppression can be halted by drugs that target immune checkpoints of the innate and antigen-specific host response. Unfortunately, not only do these approaches have limitations that prevent their use with the majority of cancer patients but, with some exceptions, predicting patient response to either class of drugs remains challenging.

Indeed, multiple studies have shown that only 5–11% percent of cancer patients derive any clinical benefit from targeted therapy and that the effect is almost invariably short lived. Specifically, most tumors lack targetable mutations, and even when they do have such mutations, the benefits of pharmacological agents are often limited to specific cancer types or contexts. Even when patients initially respond, the outcome of both approaches is usually the emergence of a new drug-resistant form of the tumor. Thus, the ability to predict which patients would likely experience a durable response to one or more of the clinically available drugs would be extremely useful.

Califano and colleagues have shown that, within the individual tumor, cancer cells can adopt only a relatively limited, discrete, and remarkably stable repertoire of transcriptional states and that these states are equally remarkably conserved across patients with the same tumor subtype (e.g., triple negative breast cancer). These states are controlled by tightly autoregulated protein modules—called tumor checkpoints (TC)—that comprise small, yet highly conserved, sets of MR proteins that integrate the effect of mutations in their upstream pathways.

In a paper recently published online on the bioRxiv preprint server1, Califano and his co-authors in the Department of Systems Biology, as well as at Columbia University Irving Medical Center (CUIMC), the Herbert Irving Comprehensive Cancer Center (HICCC), Memorial Sloan Kettering, and Emory University, report two approaches to prioritize effective treatments for clinical translation. One is aimed at targeting individual, druggable MR proteins with their already established inhibitors (OncoTarget); the other targets the entire TC-module of a tumor, by analyzing the response—a.k.a. mechanism of action (MoA)—of tumor cells that recapitulate the patient tumor MR proteins following perturbation with ~350 clinically relevant compounds (OncoTreat). The study included patients with 18 distinct aggressive human malignancies who had aggressive tumors that had progressed on at least one standard systemic therapy, with the majority having received three or more lines of treatment.

Critically, stratification of predicted drug sensitivities across large tumor cohorts showed that patients fall into a relatively small number (two to seven) of clusters (pharmacotypes) predicted to be sensitive to the same drugs. This is important because it shows that only a limited number of drugs may be required to treat the majority of tumors in a cohort, rather than requiring a different treatment for every patient, which would be clinically unfeasible. The au-

thors found that predicting drugs based on their ability to target a single MR of the entire TC-module was highly effective, based on validation in patient-derived xenografts (transplants of the tumor in a mouse), with >90% of treated tumors not able to double their volume over the course of the study, vs. 0% of the tumors treated with randomly selected antineoplastic drugs. Critically, the study confirmed the expectation that treating mice with drugs capable of targeting the entire TC-module induced statistically significant more durable responses than treating them with drugs targeting an individual MR protein.

OncoTarget and OncoTreat have several practical advantages over current approaches. First, they are based on the analysis of RNA rather than DNA. Not only is RNA profiling much cheaper and faster but it tracks the evolution of the tumor over time, allowing long-term management of the disease. (It is rare for new druggable mutations to emerge following progression.) In addition, it can be applied to individual subpopulations of cancer cells co-existing within the same tumor mass, thus opening the road to rational combination therapy. They are also the only RNA-based tests that are both NY/CA Dept. of Health approved and CLIA compliant, thus allowing their use in a clinical context.

Second, based on the benchmarking of more than 12,000 primary tumors from The Cancer Genome Atlas (TCGA), as well as 100s of samples from patients with metastatic cancers unresponsive to treatment, OncoTarget and OncoTreat can prioritize multiple candidate treatments for virtually every patient. Though not all identified drug candidates will turn out to be effective, this provides an important starting point, especially for patients with rare tumors like the one in the case study presented in the paper, which cannot undergo the kind of cohort-based studies that have been performed for more prevalent malignancies.

Third, the presence of well-defined pharmacotypes—i.e., tumors with shared predicted drug sensitivity—in virtually all the cancer cohorts the researchers have studied, supports the prioritization and evaluation of predictions through standard basket and umbrella trials, respectively, with the OncoTarget and OncoTreat tests. This includes patients across multiple malignancies predicted to share sensitivity to the same drug or drugs. Indeed, an OncoTreat/OncoTarget-based basket study for pancreatic ductal adenocarcinoma is already under way at Columbia. Fourth, the data suggest that pharmacological targeting of TC-modules leads to more durable clinical responses than targeting individual proteins, thus providing a single-agent form of combination therapy.

Finally, the approach has the potential to capture changes not necessarily driven by new mutations, such as metastatic progression and therapy resistance, allowing the clinician to adapt the therapy to the dynamic nature of the tumor and even to co-existing subpopulations with different drug sensitivity. Regarding this last point, in light of the ability of VIPER (an algorithm developed in the Califano lab) to accurately and reproducibly measure protein activity in single cells, as shown in a recent *Cell* manuscript[2], the researchers are extending the OncoTarget and OncoTreat methodologies to the single-cell level. This will allow drug prioritization for independent subpopulations co-existing in a tumor with distinct drug sensitivities, potentially avoiding drug resistance before it leads to relapse.

N-of-1 Clinical Application: A Case Report

The researchers found their PCM framework to be uniquely suited to identify therapeutic alternatives, even for rare cancers lacking actionable mutations and standard-of-care options. In the paper, they present the case of a 14-year-old male with calcifying nested stromal epithelial tumor (CNSET), an exceptionally rare primary hepatic tumor that occurs in children and young adults; only about 40 cases have been reported in the literature.

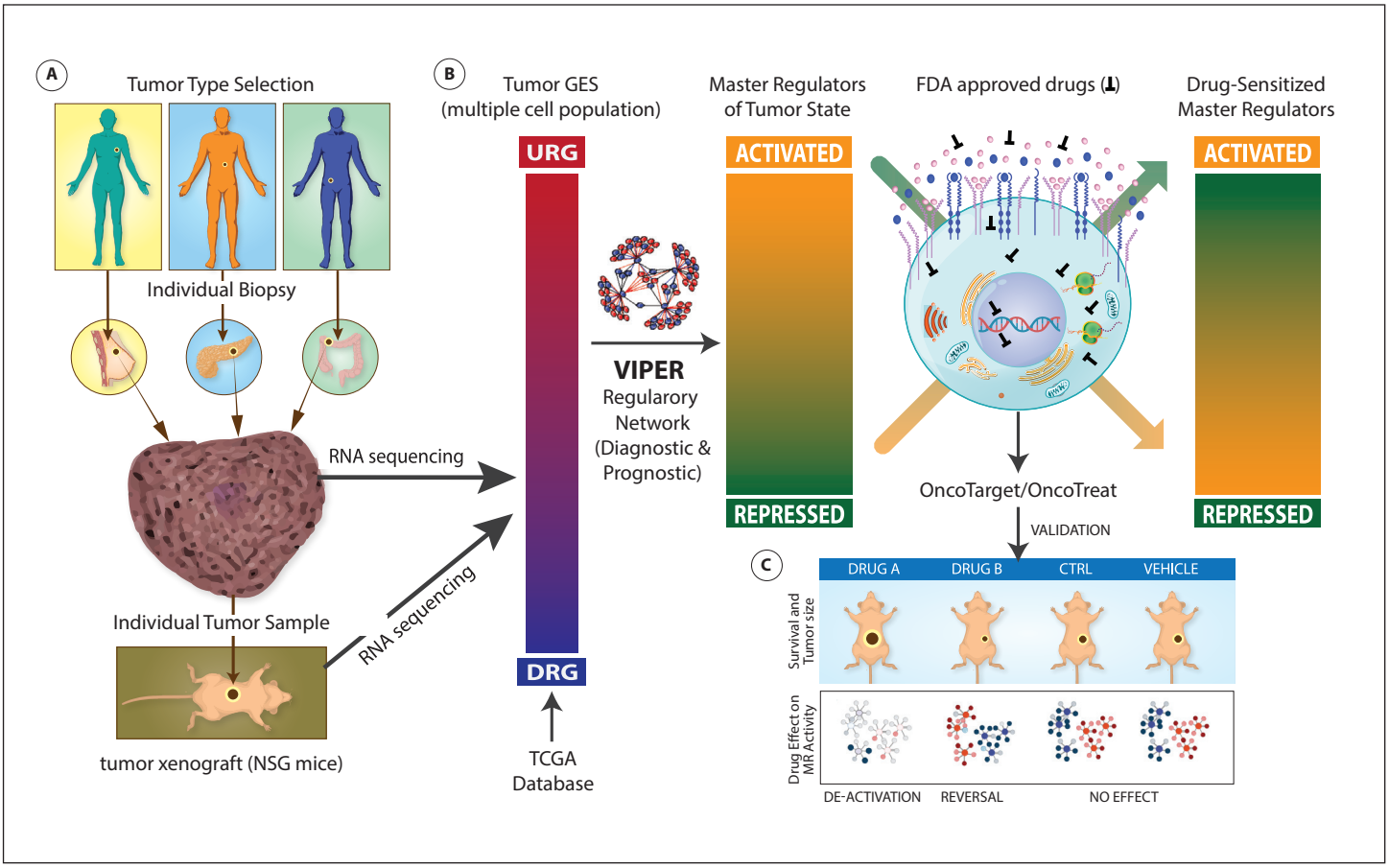
The patient had a partial response to chemotherapy and successfully underwent debulking surgery. His post-operative chemotherapy, however, was complicated by severe colitis, and his family decided to discontinue systemic therapy. Over the next six months, his disease spread to the liver and lungs. He also developed biliary obstruction and transaminitis, making him ineligible for clinical trials and precluding the use of most chemotherapy agents.

Given the lack of remaining viable therapeutic options, tumor tissue was sent for the CLIA-certified OncoTarget test. The most significantly activated targetable protein was PDGFR-B. After discussing the results with the family, and informing them of the absence of clinical data on targeting PDGFR-B in this exceedingly rare malignancy, his doctors selected sunitinib as the best candidate drug, given its high relative selectivity for PDGFR-B, accessibility, and safety. The patient had a partial response to the first (six-week) cycle of sunitinib, which deepened by the end of Cycle 3. Remarkably, the patient continues to respond and remains on sunitinib—two years after his original presentation.

"This N of 1 study," says Califano, "suggests that we do not need a different drug for each type of tumor. Instead, a relatively small number of drugs could be selected to treat a majority of patients in specific cohorts. I should note that the N of 1 can go only as far as the repertoire of currently available antineoplastic agents, virtually none of which targets MR proteins but rather proteins upstream of them. The full potential of the N of 1 approach will be realized when the number of MR inhibitors significantly increases, which is being pursued as part of several ongoing collaborations with biotech and pharmaceutical companies."

REFERENCES
1.Mundi PS, Dela Cruz FS, Grunn A, Diolaiti D, Mauguen A, Rainey AR, Guillan KC, Siddiquee A, You D, Realubit R . . . Califano A. Pre-clinical validation of an RNA-based precision oncology platform for patient-therapy alignment in a diverse set of human

N of 1 Preclinical Trial Conceptual Diagram

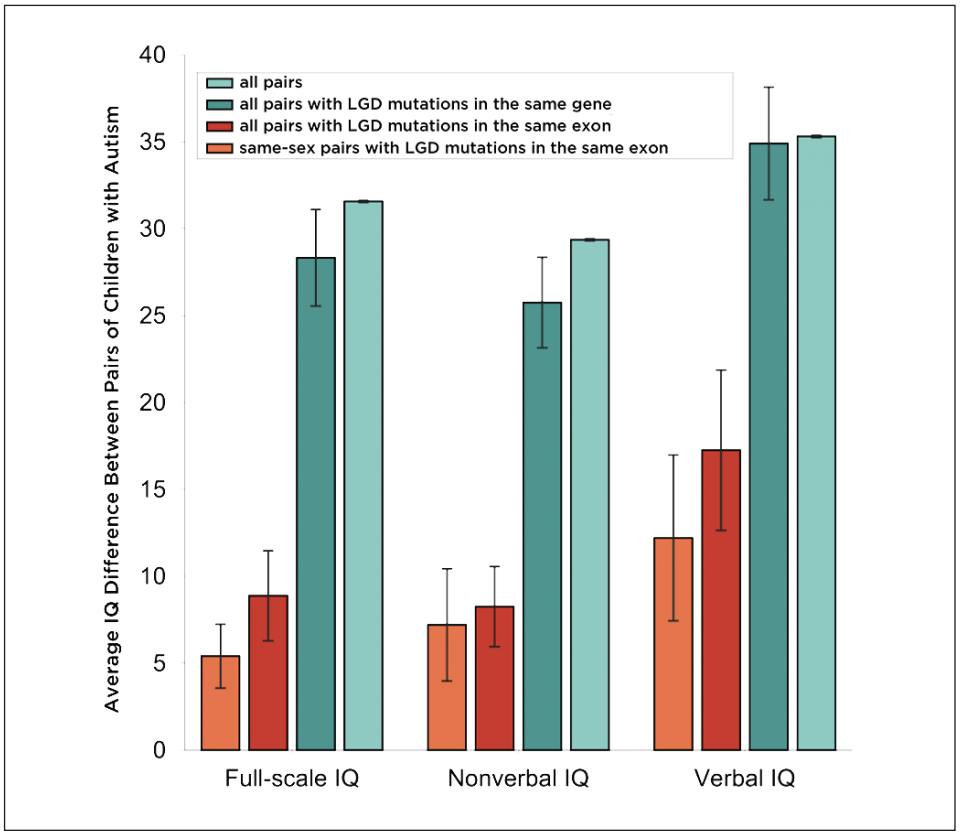


(A) Adults with metastatic solid tumors with progression or intolerance to all standard treatments and with accessible site for biopsy are enrolled. Fresh tumor tissue from biopsy is partitioned for (i) clinical histopathology review, (ii) xenografting into immunodeficient mice, and (iii) mRNA profiling (RNA-Seq). If engraftment is successful, the mature P0 passage tumor is also profiled by RNASeq to confirm candidate MR conservation between patient tumor and PDX (OncoMatch). (B) Use of VIPER, OncoTarget, and OncoTreat analysis to predict optimal drugs for PDX treatment. (i) mRNA profiles are generated from tumor samples. (ii) A gene expression signature (GES) is generated by comparing the tumor profile with a large pan-cancer RNASeq compendium (reference) comprising all TCGA samples. (iii) Cancer-type specific network(s) are used to interrogate the GES to identify the most aberrantly activated and inactivated proteins (i.e., candidate MRs) by VIPER analysis. (iv) OncoTarget identifies the most aberrantly activated proteins among those for which a high-affinity inhibitor drug is available (i.e., druggable MRs)—e.g., receptor and intracellular kinases, cell surface molecules, and enzymes involved in epigenetic regulation. (v) OncoTreat identifies the drugs inducing the strongest activity inversion of all candidate MRs (i.e., TC-module inverter drugs) by VIPER analysis of drug-perturbation profiles generated by treating context-relevant cell line models with available approved and experimental (antineoplastic) drugs. (C) Candidate drugs are prioritized based on prediction p-value, conservation of prediction based on the PDX RNASeq profile, and clinical relevance. Mice from the P1 passage are randomized into candidate drug arms, a negative control drug arm, and a vehicle control arm.

malignancies resistant to standard treatments. bioRxiv preprint. 2021 Nov 4; doi.org/10.1101/2021.10.03.462951.

2.Obradovic A, Chowdhury N, Haake SM, Ager C, Wang V, Vlahos L, Guo XV, Aggen DH, Rathmell WK, Jonasch E . . . Califano A, Drake CG. Single-cell protein activity analysis identifies recurrence-associated renal tumor macrophages. Cell 2021 May 27;184(11):2988-3005. doi: 10.1016/j.cell.2021.04.038.

Diversity and Severity of Autism Symptoms Linked to Mutation Locations



Among children with autism spectrum disorder, IQ can be substantially different even when the same gene is affected by a likely gene-disrupting (LGD) de novo mutation. When such mutations affect the same exon , however, IQs are much more similar, the new study found.

One of the most recognizable characteristics of autism is an amazing diversity of associated behavioral symptoms. Clinicians view autism as a broad spectrum of related disorders, and the origin of the disease's heterogeneity has puzzled scientists, doctors, and affected families for decades.

In a recent study, researchers at Columbia University Vagelos College of Physicians and Surgeons have made an important step towards understanding the biological mechanisms underlying the cognitive and behavioral diversity of autism cases triggered by de novo truncating mutations. These mutations occur in parents' germline cells and usually strongly disrupt the functions of target genes. De novo truncating mutations are responsible for close to 5% of autism cases and up to 20% of cases seen clinically.

Autism spectrum disorders that are triggered by a single disrupted gene represent a relatively simple genetic type of the disease. The perplexing observation that scientists were grappling with for many years is that even when truncating mutations occur in the same gene, they often lead to a wide range of symptoms and behavioral patterns in different children.

> "It turns out that we weren't looking closely enough at how and where an autism gene is mutated."
>
> —Dennis Vitkup,

The new study found that the severity of autism symptoms often depends on which specific functional unit within a gene is the target of a mutation.

"It turns out that we weren't looking closely enough at how and where an autism gene is mutated," says study leader Dennis Vitkup, PhD, associate professor of systems biology and of biomedical informatics at Columbia University Vagelos College of Physicians.

Human genes, similar to genes of other eukaryotic species, are composed of separate coding units, called exons, which are frequently joined together in different combinations across tissues and developmental stages. "Upon closer examination, we found that different children with truncating mutations in the same exon have strikingly similar behavioral symptoms and disabilities," Vitkup says.

The study was published online in the journal *Molecular Psychiatry.*

Same exon, similar symptoms

In the study, Vitkup and colleagues Andrew H. Chiang, Jonathan Chang, and Jiayao Wang, analyzed genetic and clinical data from over 2,500 people with autism, focusing on cases resulting from truncating mutations.

Among children with autism spectrum disorder, IQ can be substantially different even when the same gene is affected by a likely gene-disrupting (LGD) de novo mutation. When such mutations affect the same exon, however, IQs are much more similar, the new study found. When the researchers compared random pairs of children with autism, they found that their nonverbal, verbal, and overall IQ scores differed on average by more than 30 points. Children with truncating mutations in the same gene showed similar differences.

However, when the researchers compared autistic children affected by mutations in the same exon of the same gene, their IQs differed by less than ten points, which is comparable to the IQ measurement errors. The researchers observed very similar patterns for multiple other scores characterizing children's communication, social, and motors skills.

"This tells us that, with autism-associated truncating mutations, it's the exon, and not the whole gene, that often represents a functional unit of impact," Vitkup says.

More severe symptoms associated with frequently used exons

The researchers demonstrated that the behavioral and cognitive severity of autism is proportional to the likelihood with which targeted exons are used in gene transcripts, with more severe effects associated with mutations in more frequently used exons. When mutations occur in the same exon, the resulting expression-level changes are especially similar, leading to similar clinical consequences.

Surprisingly, the study also showed that the gene expression changes caused by truncating mutations can be quite mild. "Our analysis demonstrates that autism cases can be triggered by relatively small changes in overall gene dosage, often as small as 15%," says the study's first author Andrew Chiang, a graduate student in the Department of Biomedical Informatics.

Implications for precision medicine

The study may have significant implications for precision medicine. Diagnostic and prognostic tests may now pay special attention to specific exons affected by truncating mutations.

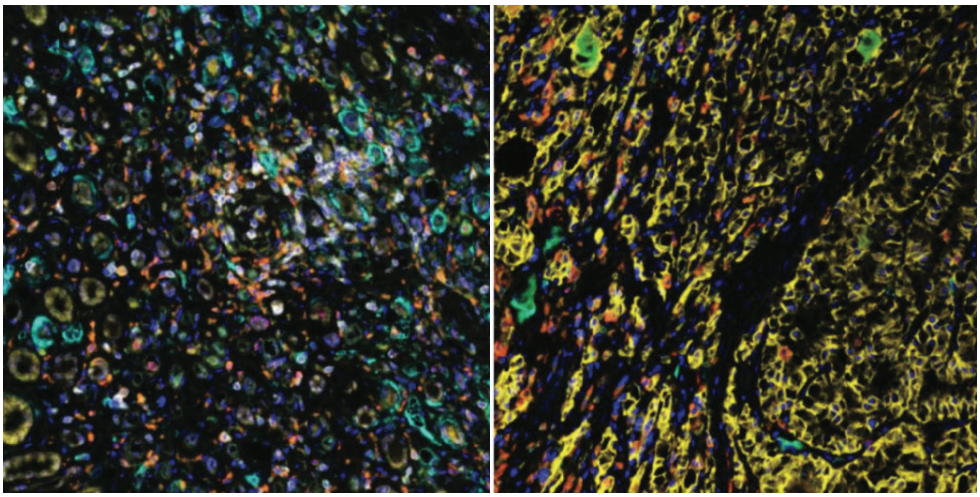
The study also suggests a therapeutic approach for alleviating the consequences of truncating mutations in autism. "It would be very hard to develop drugs for thousands of different mutations in many hundreds of target autism genes," Vitkup says, "but our study demonstrates that behavioral abnormalities often originate from relatively small decreases in the target gene's dosage. These genetic insults may be, at least partially, compensated by increasing the expression of an unaffected gene copy using new molecular tools such as CRISPR."

—Reprinted with permission by Columbia News

REFERENCE

Chiang AH, Chang J, Wang J, Vitkup D. Exons as units of phenotypic impact for truncating mutations in autism. Molecular Psychiatry 2021 May;26(5):1685-1695. doi: 10.1038/s41380-020-00876-3.

New Single-Cell Analysis Tool Links Immune Cells to Kidney Cancer Recurrence



Normal kidney tissue (left) and typical tumor tissue (right). Images from Obradovic et al. (2021).

The immune nature of kidney cancer stands out when compared to other cancers: More immune cells infiltrate kidney cancers than most other solid tumors, and kidney cancer is one of the most responsive malignancies to today's immunotherapy regimens.

But despite treatment, many patients with clear cell renal carcinoma—the most common type of kidney cancer—eventually relapse and develop incurable metastatic disease.

A new study shows that the presence of a rare and previously unknown type of immune cell in kidney tumors can predict which patients are likely to have cancer recur after surgery. These cells could even be driving aggressive disease.

"Our findings suggest that the presence of these cells could be used to identify patients at high risk of disease recurrence after surgery who may be candidates for more aggressive therapy," says co-senior author Charles Drake, MD, PhD, adjunct professor of medicine at Columbia University Vagelos College of Physicians and Surgeons and the Herbert Irving Comprehensive Cancer Center.

The study was published online May 20 in the journal *Cell*. Andrea Califano, Dr, the Clyde and Helen Wu Professor of Chemical and Systems Biology and chair of systems biology at Columbia University Vagelos College of Physicians and Surgeons, also is co-senior author of the study.

New tools tapped to profile cells

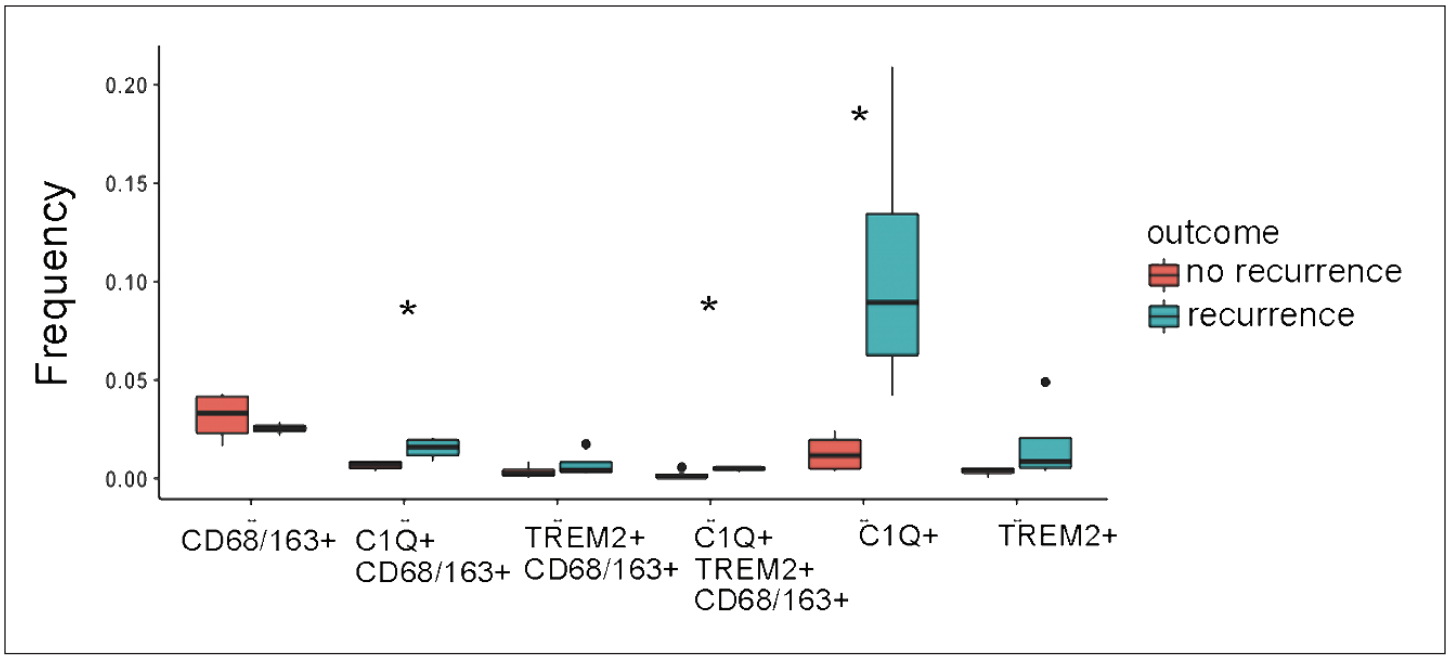
Though kidney tumors are densely infiltrated by immune cells, cell subtypes and their association with post-surgical outcomes have remained largely unknown.

It's like looking down at Manhattan and seeing that large numbers of people from all over travel into the city every morning, says Aleksandar Obradovic, an MD/PhD student at Columbia University Vagelos College of Physicians and Surgeons and the study's co-first author. "To understand how these diverse commuters are interacting with Manhattan residents, we need finer details: Who are they; what are they like, where do they go, and what are they doing?"

To uncover the fine details of the immune cells that infiltrate kidney cancers, the researchers combined two of the newest techniques in cancer research.

The first, called single-cell RNA sequencing, captures a snapshot of gene activity in individual cells within a tumor. This high-throughput technique allows researchers to obtain such snapshots inside of tens of thousands of cells from one tumor in a single experiment, providing insights into the identity and behavior of the various cell types.

This powerful technique can identify new types of cells, but there is a drawback. Because single-cell sequencing works by detecting a small number of



C1Q+ macrophages in the tumor stroma of patients with kidney cancer is related to recurrence.

mRNA molecules inside each cell, it often fails to detect the mRNAs of genes with low expression levels, including key signaling genes and drug targets such as immunotherapy checkpoints.

"In many experiments, single-cell RNA sequencing misses up to 90% of gene activity, a phenomenon known as gene dropout," Obradovic says.

Prediction algorithm addresses gene dropout

The researchers addressed gene dropout by developing a prediction algorithm that can infer which genes are active by looking at the expression of other related genes. "Even when a lot of the data are missing due to dropout, we still have enough clues to infer the activity of the upstream regulator gene," Obradovic says. "It's like playing 'Wheel of Fortune': I can usually guess what's on the board even when most of the letters are missing."

The algorithm, called meta-VIPER, builds on the VIPER algorithm developed in the Califano laboratory.

With the addition of metaVIPER, the researchers estimate they can accurately detect the activity of 70% to 80% of all regulatory genes in each cell, eliminating dropout across cells.

Patient outcomes track with newly identified macrophages

This combined approach was used to analyze more than 200,000 tumor cells and normal cells in adjacent tissue taken from 11 patients with clear cell renal carcinoma who underwent surgery in the Department of Urology at Columbia.

C1Q+ macrophages in the tumor stroma of patients with kidney cancer is related to recurrence.

The analysis revealed a unique sub-population of immune cells called macrophages found only in tumors and associated with eventual relapse of disease after initial treatment. The VIPER analysis also revealed the top genes (or master regulators) that control the activity of these macrophages. This "signature" was validated in a second set of patient data obtained through a collaboration with researchers from Vanderbilt University; here the signature strongly predicted relapse in a second set of over 150 patients.

Furthermore, these macrophages were found to interact directly with tumor cells through receptor-ligand gene pairs. "These data raise the intriguing possibility that these macrophages are not just markers of more risky disease but may actually cause the disease to recur and progress," Obradovic says, "and that targeting these cells could improve clinical outcomes."

Thus VIPER-based technologies, such as the Oncotreat test, could be used to identify drugs targeting these rare but critical subpopulations, thus preventing the poor outcomes associated with their presence, Califano says.

Techniques could be applied to other cancers and diseases

The combination of single-cell sequencing with the VIPER algorithm has potential to dissect other types of cancer too, the researchers say.

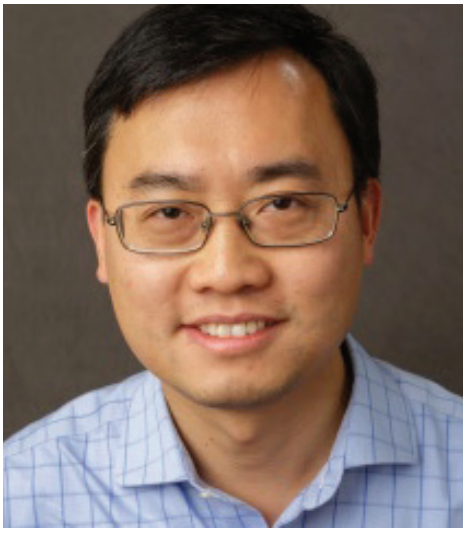
"Our study demonstrates that the two techniques, when combined, are highly effective at characterizing the cells within a tumor and in surrounding tissues and should have broad applicability, even beyond the study of cancer," Drake says.

—Reprinted with permission from Columbia News

REFERENCE

Obradovic A, Chowdhury N, Haake SM . . . McKiernan J, Califano A, Drake CG. Single-cell protein activity analysis identifies recurrence-associated renal tumor macrophages. Cell 2021 May 20;184(11):P2988-3005. doi.org/10.1016/j.cell.2021.04.038.

FACULTY SPOTLIGHT:
Q&A with Yufeng Shen



Yufeng Shen is an associate professor in the Department of Systems Biology and the Department of Biomedical Informatics. He received his BSc in biochemistry and molecular biology from Peking University (where he eventually realized that he was more at home in front of a computer than in a wet lab) and his PhD in computational biology from Baylor College of Medicine. He came to Columbia in 2008 as a postdoc in Computer Science and joined the faculty in Systems Biology and Biomedical Informatics in 2011.

Q: *Why did you decide to get your PhD in computational biology? And please touch upon James Watson's genome.*

A: I had always felt at home in physics and mathematics, yet at the same time I was intrigued to understand how genes affect human traits and diseases. The field of computational biology seemed the perfect place to delve into mathematics and physics while helping to make headway in understanding disease—which, in addition to intellectual satisfaction, brings the gratification of directly affecting people's lives.

The field of computational biology has advanced considerably, just in the time since I earned my PhD. A lot of that progress is directly attributable, of course, to the tremendous gains in computing power. It's very exciting when the technology is there to help you answer the questions you want to answer.

While at Baylor College of Medicine, I participated in the sequencing of James Watson's genome. Watson, who shared a Nobel Prize for describing the double-helix structure of DNA, was the second person to publish his fully sequenced genome. Using massively parallel DNA sequencing, we were able to complete the sequencing in two months, at one-hundredth the cost of earlier sequencing methods.

The significance of the task was not the technical feat of full-genome sequencing in itself—impressive as that was—but the promise it held for human health. The association of genetic variation with disease and drug response, specifically, holds great promise for the uses of "genomic medicine."

Q: *Please describe your work in genetic analysis.*

A: A major focus of my lab's work is to link genetic variants to human conditions. This occupies about 40 percent of our current resources and efforts. We're motivated by basic genetic questions, even though in the long term, findings can have meaningful clinical application. While it is critical to identify which variants cause which disease, not every person carrying a specific variant will develop that disease. Sometimes it's because it's a recessive gene, and an individual carries only one copy. But often the picture is more complex—and we don't fully understand the many biological, and perhaps even environmental, factors that influence whether a disease manifests. Genetic architecture is very complex, and most diseases cannot be explained by a variant at a single location.

Q: *What hereditary diseases in particular are you studying?*

A: Right now, we're focusing on early-onset conditions such as congenital heart disease and autism. About 3 percent of newborns have congenital heart disease. In about two thirds of cases the condition is relatively benign. But in the remaining one third of cases, it can be severe, with a high rate of mortality and issues affecting development and long-term heart function. If we had a better understanding of the disease mechanism, identifying the variant in a newborn who showed signs of a severe condition could enable a definitive diagnosis, and perhaps inform effective medical intervention.

We're also working with the Simons Foundation on its SPARK autism research initiative. What's unique about the project is its innovative online recruitment approach and its focus on ongoing communication with the participating families. Any parent can enroll, and they already have about 100,000 families, which is a huge number. We've analyzed about 20,000 so far and have identified several new risk genes. The hope is that with the full study we will find nearly all the high-impact risk genes, providing a foundation for deeper understanding of the condition and early interventions.

In addition to congenital heart disease and autism, we've been working on congenital diaphragmatic hernia, pulmonary hypertension, tracheoesophageal defects, and breast cancer.

Q: *How does your work in genetic analysis tie in to your work developing new computational methods?*

A: I would say another 40 percent of our current work is focused on developing new computational methods for what is called computational genomics. This work is symbiotic with the search for genetic causes of human conditions. The computational methods enable the search, and the search validates the methods and points the way toward further computational advances. A particular question we try to answer is how do we interpret the functional impact of a

genetic mutation. We ask how mutations change protein function. A missense mutation is one that produces an amino acid different from the usual one. About 20 to 30 percent of these mutations have a functional impact.

This information cannot be inferred solely from biophysical modeling; we combine the modeling with machine learning enabled with large data sets. If a mutation causes damage, then in a healthy population it should be less frequent than would be expected by chance. We also look at other species. If a mutation has a functional impact, its position is likely conserved across species. Our latest computational methods use deep learning to model protein sequences.

Q: *Please tell us about your work in computational immunology.*

A: Currently, about 20 percent of the lab's time is spent directly on immunological questions. We use computational analysis and mathematical modeling to understand how the immune cells work, particularly in humans. One interesting question—especially relevant in the time of COVID-19—is the role of T-cells in the body's immunological response. Specifically, we are looking at how T-cells recognize antigens. The star of the vaccine response is B cells that produce antibodies, but T-cells are the hidden helpers and coordinators of B cell response. We are interested in predicting what kind of T cells can recognize a particular antigen, such as a fragment from a protein of a virus or our own body, as well as what kind of antigens a particular T cell can recognize. This is a very challenging computational problem. Solving it would open up a new way to ask questions in many research areas such as vaccine design, cancer immunotherapy, autoimmunity, and organ transplantation.

Q: *Is there anything you'd like to add?*

A: Many academics complain about such things as the crushing funding pressure, long working hours, etc. These are all true, but I really like my job. I particularly enjoy working with students and postdocs and seeing them mature as scientists. I also have top-notch collaborators and colleagues, who are not only excellent scientists or physicians, but kind human beings. Overall, I feel privileged to be have been in the department for 10 years, and I'm excited about the decades to come.

Around the Department, 2020-2021

Selected Grants and Awards

Systems Biology will receive $3,829,859 over five years from the National Institute of Mental Health for "Discovery and Analysis of Brain Circuits and Cell Types Affected in Autism and Schizophrenia." The project will be led by **Dennis Vitkup,** PhD, and **Joseph Gogos,** MD, PhD.

Corinne Abate-Shen, PhD, Molecular Pharmacology & Therapeutics, will receive $2,369,258 over five years from the National Cancer Institute for "Molecular Mechanisms of Prostate Cancer Metastasis."

Barry Honig, PhD, Systems Biology, will receive $2,025,000 over four years from the National Institute of General Medical Sciences for "Genome-Wide Structure-Based Analysis of Protein-Protein Interactions and Networks."

Tal Korem, PhD, Systems Biology and Obstetrics & Gynecology, will receive $3,404,285 over five years from the Eunice Kennedy Shriver National Institute of Child Health and Human Development for "A large scale investigation of the vaginal metagenome and metabolome and their role in spontaneous preterm birth." Korem also will receive $307,136 over five years from the National Institute of Nursing Research for a subaward of "The Role of Host-Microbial Interactions in Altering Preterm Birth Risk Among Black Women."

Melissa McKenzie, PhD, a second-year postdoctoral research scientist in the lab of Chaolin Zhang, PhD, has been awarded a K99/R00 "Pathway to Independence" award.

Raul Rabadan, PhD, will receive $6,804,000 over seven years from the National Cancer Institute for the project "Towards a quantitative understanding of tumor evolution."

Guillaume Urtecho, PhD, a post-doctoral fellow in Harris Wang's lab, has been named by the Howard Hughes Medical Institute as a 2020 Hanna Gray Fellow.

Harris Wang, PhD, received the Vilcek Prize for Creative Promise in Biomedical Science. Wang also will receive $629,997 over two years from the Department of Energy for a subaward of "Secure Biosystems from Sequence to Cell to Populations" and $273,648 over five years from the National Science Foundation for a subaward of "The Rules of Microbiota Colonization of the Mammalian Gut."

Harris Wang, PhD, will receive $2,665,170 over four years from the National Institute of Biomedical Imaging and Bioengineering for "A high-performance and versatile technology for precision microbiome engineering" project.

Xuebing Wu, PhD, has been awarded the 2021 Pershing Square Sohn Prize for Young Investigators in Cancer Research for his innovative approaches to cancer research.

Chaolin Zhang, PhD, will receive $2,066,788 over four years from the National Institute of General Medical Sciences for the project "Integrative analysis of tissue-specific alternative splicing regulation under adaptive selection."

Three early-career scientists at VP&S—**X. Shawn Liu** (Physiology & Cellular Biophysics), **Xuebing Wu** (Medicine), and **Nikhil Sharma** (Molecular Pharmacology & Therapeutics)—have been selected as 2021 Paul A. Marks Scholars.

Two Systems Biology GRAs receive awards:

Tomasz Blazejewski: Dean's Award for Excellence in Research

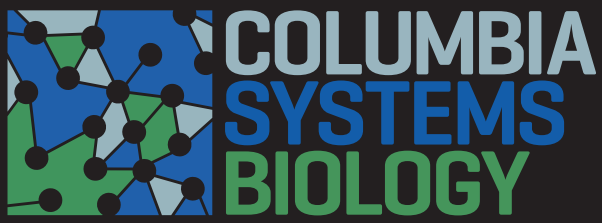
Hanna Mendes Levitin: Titus M Coan Award for Excellence in Research

PHD GRADUATES

Congratulations to our recent grads!

Siying Chen (Shen lab)
Elise Flynn (Lappalainen lab)
Benjamin Hobson (Sims lab)
Sunny Jones (Califano lab)

COLUMBIA UNIVERSITY
IRVING MEDICAL CENTER



Columbia University Department of Systems Biology
Irving Cancer Research Center
1130 St. Nicholas Avenue
New York, NY 10032

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