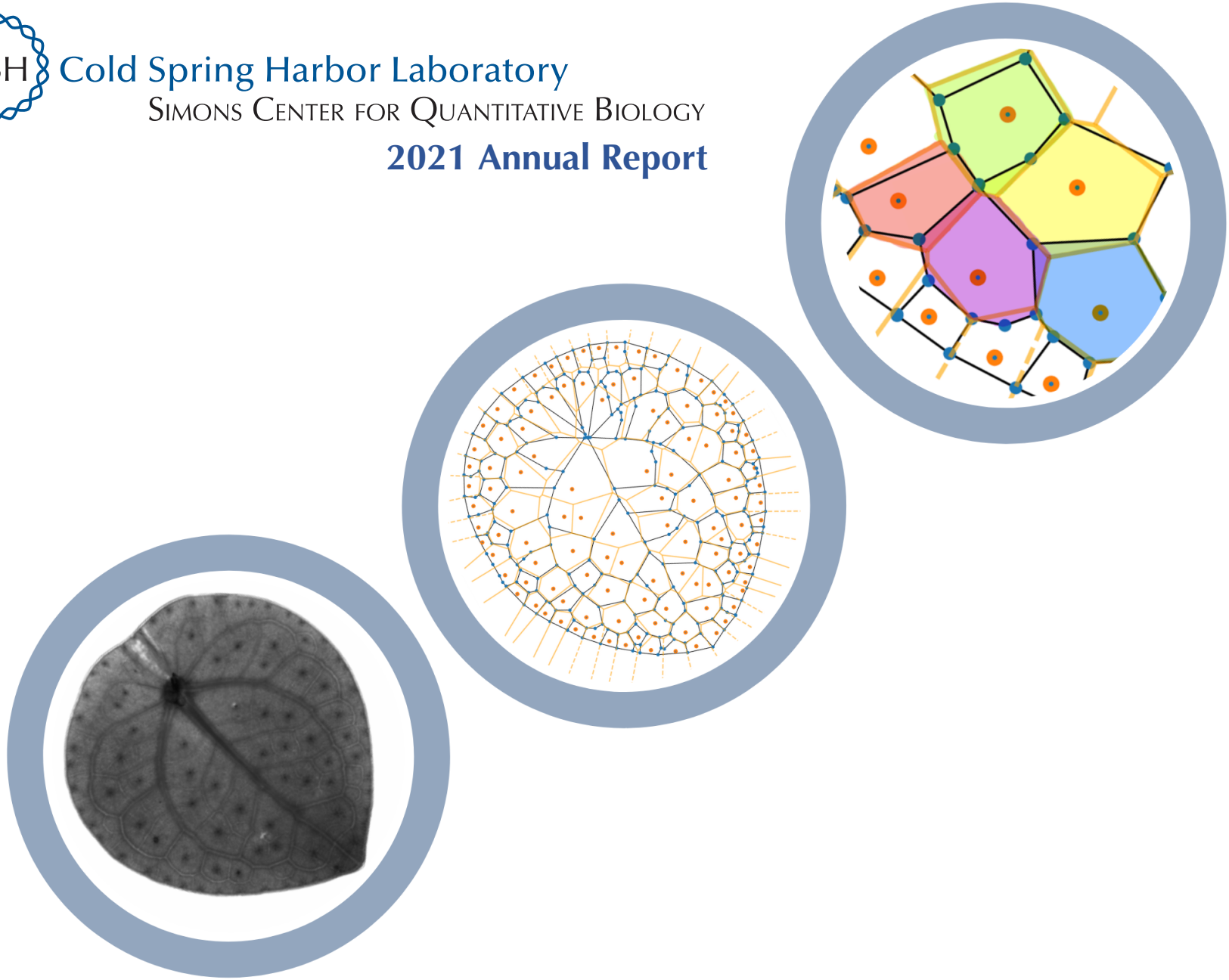
Cold Spring Harbor Laboratory

Simons Center for Quantitative Biology

2021 Annual Report



From left to right, brightfield microscopy image of a branch of a tomato plant, manual tracing of veins and hydathodes, algorithmic analysis of venation patterning. Images courtesy of Ph.D. student Xingyu (CiCi) Zheng.

LETTER FROM THE CHAIR

AS IT HAS BEEN EVERYWHERE, 2021 at the SCQB has been a roller coaster of a year. We soared with the vaccines and then sank with new SARS-CoV-2 variants. We rose as we returned to full-time work on campus, and dipped with the re-introduction of social distancing. We went up as the masks came off, and returned to earth as we regretfully applied them again.

It has been a great relief to resume most face-to-face interactions on campus, but travel remains limited and scientific meetings have almost all been virtual. Overall, it has been a good time to think and read deeply; to write computer code, work with equations, and analyze data; and to clear backlogs of scientific writing.

Accordingly, we produced 40 new preprints and publications this year, including a record 12 original manuscripts posted as preprints, as well as a number of high-impact journal publications (p. 10). New grants were down somewhat, at $2.8M, but this figure does not include several recent NIH submissions that scored exceptionally well in review and are likely to be funded in the coming months. It also does not include major grants awarded to associated members of the center, such as Tatiana Engel's NIH BRAIN Initiative grant (p. 9).

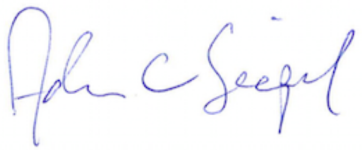
The center continues to grow, as the newer research groups obtain funding and staff up. The total size of all ten member research groups is now 63, including faculty, staff, and trainees, despite a pandemic-related slow-down in the recruitment of postdoctoral associates that is only now beginning to relax. Perhaps also because of the pandemic, relatively few students and postdocs moved on to new positions this year. Nevertheless, we were pleased to grant Ngoc (Tumi) Buu Tran her Ph.D. (p. 8), and delighted to have Drs. Wei-Chia Chen and Juannan Zhou successfully obtain tenure-track faculty positions (p. 8).

One unforeseen benefit of the pandemic is that virtual scientific meetings have blossomed, in part owing to steady improvements in technology and meeting organization. Members of the SCQB organized five major virtual meetings this year, covering topics ranging from machine learning to algorithms to molecular evolution (p. 7). These meetings have been a vital way for scientific communication to continue as travel remains severely curtailed.

In other developments, several QB members participated in the successful renewal of the CSHL Cancer Center grant, which now features QB as a major theme and includes several of our faculty members. We held a productive meeting of the External Advisory Committee (p. 14) for the SCQB in May. We launched a new "Mentored Independent Study" program for Ph.D. students hoping to strengthen their academic training in quantitative disciplines not well supported by existing courses at CSHL (p. 8). We participated in the search for a new Chief Information Officer at CSHL, which resulted in the recent hire of Doug Torre. We granted a new "Interdisciplinary Scholar" award to Hannah Meyer and Saket Navlakha, who are recruiting a joint postdoctoral fellow for their project (p. 7). And we congratulated Hannah Meyer for being named the "Early Career Researcher of the Year" by the UK Biobank (p. 9).

As we spend a second consecutive holiday season sheltering from this unremitting virus, I remain hopeful that we are on the threshold of better times for scientific research. With continued progress in vaccination, testing, and treatment, it might be possible in 2022 to resume some travel and in-person scientific meetings, with appropriate cautionary measures. May next year's letter tell a happier story.

Best Wishes for the New Year,



Adam Siepel, Ph.D., Chair
December 29, 2021

OVERVIEW

THE SIMONS CENTER FOR QUANTITATIVE BIOLOGY IS focused broadly on revealing how genomes work, how they evolve, and what makes them go wrong in disease. Investigators at the SCQB pursue diverse research interests in a wide variety of areas, but our research continues to be permeated by four major themes: **Gene Regulation, Evolutionary Genomics, Genomic Disease Research, and Genomic Technology Development.**

Our Faculty

Molly Gale Hammell
Associate Professor
Cancer Center Member
Gene regulation, Genomic disease research

Ivan Iossifov
Associate Professor
Evolutionary genomics, Genomic disease research, Gene technology development

Justin Kinney
Associate Professor
Cancer Center Member
Gene regulation, Genomic technology development

Peter Koo
Assistant Professor
Cancer Center Member
Gene regulation

Alexander Krasnitz
Associate Professor
Cancer Center Member
Evolutionary genomics, Genomic disease research

Dan Levy
Associate Professor
Cancer Center Member
Genomic disease research, Genomic technology development

David McCandlish
Assistant Professor
Cancer Center Member
Gene regulation, evolutionary genomics

Hannah Meyer
CSHL Fellow
Cancer Center Member
Gene regulation, Genomic disease research

Saket Navlakha
Associate Professor
Evolutionary genomics

Adam Siepel
Professor
Cancer Center Program Co-Leader
Gene regulation, Evolutionary genomics

GENE REGULATION

Researchers at the SCQB develop theoretical and experimental methods, together with new computational tools, for elucidating the relationship between biological sequences and biological functions ranging from gene expression to protein function. Ongoing studies in this area address the determinants of transcription elongation rates and RNA stability, the code for RNA splicing, the behavior of small non-coding RNAs, and the impact of transposable elements on gene expression. In addition, researchers at the SCQB are broadly interested in mathematical modeling of the regulation of gene expression in mammals, ranging from transcription factor binding and chromatin accessibility, to transcription initiation and elongation, to the determination of RNA stability.

EVOLUTIONARY GENOMICS

Scientists at the SCQB develop theory and mathematics to address several open questions in evolutionary genetics, including the processes by which gene expression evolves in mammals, the genetics of speciation, and the evolutionary implications of mutations that interact to determine organismal fitness. Additional studies at the Center use evolutionary methods to identify regulatory elements, reconstruct early human history, and estimate the fitness consequences of new mutations in the human genome. Researchers also use evolutionary signatures to identify genes associated with autism spectrum disorder and employ phylogenetic methods to study the evolution of tumors.

GENOMIC DISEASE RESEARCH

Researchers at the Center are trying to understand the genetics of autism spectrum disorder (ASD) by analyzing large genomic data sets. Others are developing mathematical and statistical tools to characterize the cellular composition, genomic disruptions, evolutionary history, and invasive capacity of malignant tumors. Scientists at the Center are also investigating the role of transposable element activation in neurodegenerative diseases, particularly amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD).

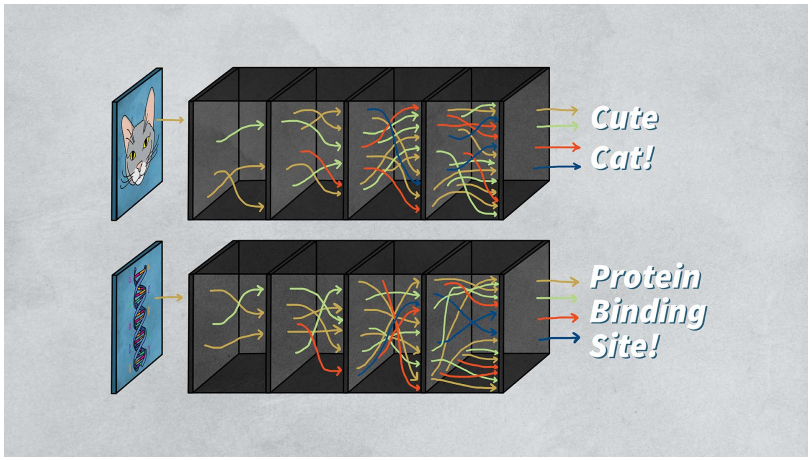
Researchers are also interested in diverse modeling and statistical inference problems in cancer and immunology, often using single-cell sequencing data. Most SCQB faculty members are active participants in the CSHL Cancer Center. Two of the three major Cancer Center programs include our faculty: Cancer Genetics and Genomics (CGG) and Gene Regulation and Cell Proliferation (GRCP).

GENOMIC TECHNOLOGY DEVELOPMENT

Research groups in the SCQB are developing new DNA and RNA sequencing methods, single-cell genomic technologies, and cancer diagnostics. Our scientists have also pioneered the development of massively parallel reporter assays for characterizing the relationship between regulatory sequences and gene expression, including both transcription and RNA splicing.

IN 2021, MEMBERS OF THE SCQB produced publications and preprints describing a wide variety of research projects, as summarized on p. 4-6. Below we highlight several studies of interest.

AI researchers ask: What's going on in the black box?



"Black box" demonstrates how researchers can train artificial brain-like neural networks to classify images—illustration by Ben Wigler.

Brain-like artificial networks are often referred to as a "black box" because researchers do not know how they learn and make predictions to solve artificial intelligence (AI) problems. While machine-learning researchers can train a brain-like "neural net" computer to recognize objects by showing the machine many images and seeing if it classifies it correctly, they have a problem applying this technology to analyzing DNA patterns.

Assistant Professor **Peter Koo** and collaborator Matt Ploenzke of Harvard University reported a way to train machines to predict the function of DNA sequences. Koo and his team fed DNA (genomic) sequences into a specific neural network called a convolutional neural network (CNN), which resembles the real neural networks animal brains use to process images. As reported in the journal *Nature Machine Intelligence*, Koo and Ploenzke introduced a new method to teach important DNA patterns to one layer of their CNN, allowing the neural network to build on the data to identify some key features that lead to the computer's decision-making process.

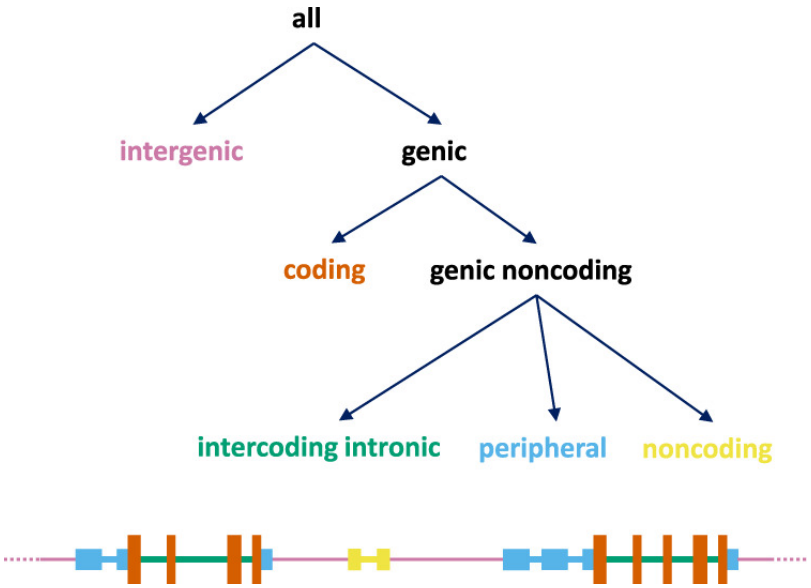
In the past, AI researchers have improved either a neural network's interpretability or robustness. Koo aims to bridge the two, believing he will get better interpretability of his models if they are robust. Ultimately, he and his team hope that if a machine can find robust and interpretable DNA patterns related to gene regulation, it will help geneticists understand how mutations affect cancer and other diseases.

Peter Koo & Matt Ploenzke, "Improving representations of genomic sequence motifs in convolutional networks with exponential activations" was published in *Nature Machine Intelligence* on February 8, 2021.

Untangling the role of spontaneous mutations in autism

Spontaneous genetic mutations contribute to autism in 30 to 39 percent of people with the condition, an increase over previous estimates, according to a new study led by Associate Professor **Ivan Iossifov** analyzing data from the Simons Simplex Collection (SSC). They found the rate is even higher 52 to 67 percent among autistic children whose siblings do not have the condition. These spontaneous, or '*de novo*' mutations are not found in either an autistic person's parents and occur more often in autistic people than in non-autistic people. Iossifov and his colleagues, including Associate Professor **Dan Levy**, say that they may contribute primarily to autism in so-called 'simplex' families with one autistic child. Inherited mutations, by contrast, explain most cases in 'multiplex' families, which have more than one autistic child.

Some studies have identified autism-related *de novo* mutations in children from multiplex families. Still, this new work estimates that mutations play a role in only about 10 percent of these children. According to Iossifov, the discrepancy with past estimates could have to do with sampling differences that muddy comparisons. Some analyses of multiplex families have relied on cell lines that can accumulate *de novo* mutations as they grow in culture dishes a problem known as 'genetic drift' and thus skew results. Iossifov's work screens samples for those stray mutations, making it possible to distinguish the *de novo* mutations that contribute to autism.



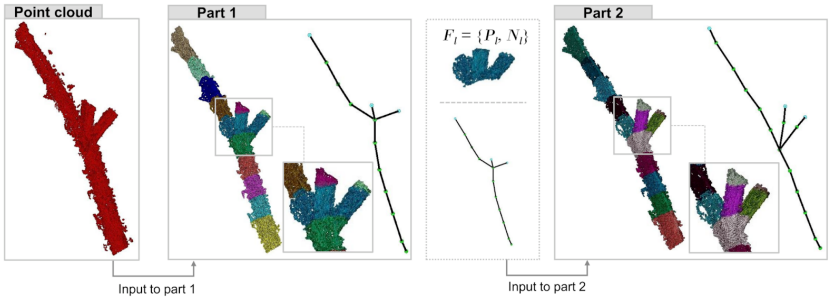
Partitioning DNA sequences in simplex and multiplex autism.

Seungtai Yoon, Adriana Munoz, Boris Yamrom, Yoon-Ha Lee, Peter Andrews, Steven Marks, Zihua Wang, Catherine Reeves, Lara Winterkorn, Abba M Krieger, Andreas Buja, Kith Pradhan, Michael Ronemus, Kristin K Baldwin, Dan Levy, Michael Wigler, Ivan Iossifov, "Rates of contributory de novo mutations in low-risk autism families" was published in *Communications Biology* on September 1, 2021.

Could machine learning improve crop production?

Machine-learning algorithms are bringing plant phenotyping into the modern age by allowing scientists to measure detailed characteristics of a plant's architecture to engineer more resilient plants or boost crop production. The advent of high-throughput 3D imaging platforms has generated an abundance of three-dimensional structures that portray the fine details of the plant's surface. Still, quantitative analysis can be challenging since the technology is so new and the datasets are so large.

Associate Professor **Saket Navlakha** teaches computer systems to analyze these three-dimensional plant structures. Navlakha has set his sights on building better plant phenotyping tools robust to noise and missing data while efficiently processing large numbers of plants. He fed high-resolution 3D images of tomato and tobacco plants into a new machine-learning algorithm to predict the plant's basic geometry and shape. Compared to existing methods for 3D skeletonization, he and his team found that their method efficiently and more accurately estimated the plant's morphology, even in areas with noisy, missing, or non-uniformly sampled data. Navlakha hopes this tool will speed up plant research by making high-throughput phenotyping faster and easier, increasing crop production and carbon sequestration.



Overview of a machine learning algorithm for predicting plant morphology.

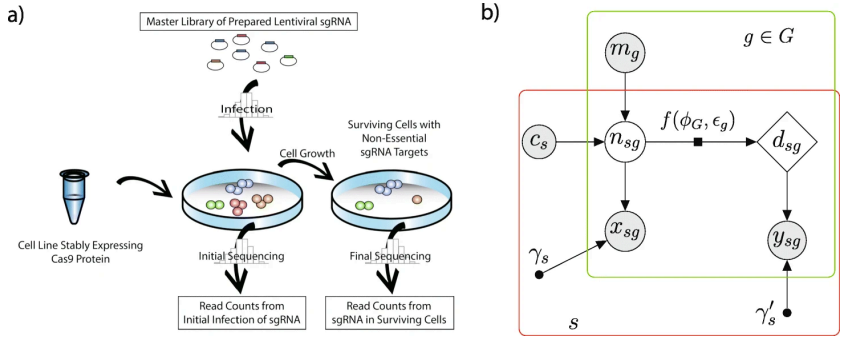
Illia Ziamtosov, Kian Faizi, Saket Navlakha, "Branch-Pipe: Improving Graph Skeletonization around Branch Points in 3D Point Clouds" was published in *Remote Sensing* on July 22, 2021.

Modeling genes essential for cancer

Elizabeth Hutton, a former Ph.D. student in the Cold Spring Harbor School of Biological Sciences, conducted her graduate studies in Professor **Adam Siepel**'s lab, finishing in 2020. Hutton's dissertation focused on the development of new statistical methods for analyzing data produced by genome-wide CRISPR screens. For a recent paper in *Genome Biology* based on her dissertation work, Hutton and Siepel teamed up with Professor Christopher Vakoc, who regularly uses CRISPR/Cas9 technology to probe the epigenetic regulation of cancer and identify new cancer drug targets.

Under the mentorship of Siepel and Vakoc, Hutton developed a mathematically rigorous modeling framework to identify genes essential for cellular growth in

particular genetic backgrounds. Her method, called Analysis of CRISPR-based Essentiality (ACE), accounts for several sources of experimental variation in CRISPR-Cas9 screens and enables new statistical tests for both absolute and differential essentiality of genes. When applied to real data, it identified a number of previously known and novel candidates for genotype-specific essentiality, including RNA m6-A methyltransferases that exhibit enhanced essentiality in the presence of inactivating TP53 mutations. This test for differential essentiality can uncover unique genetic dependencies that might otherwise be overlooked, helping to enable subtype-specific personalized cancer therapies.



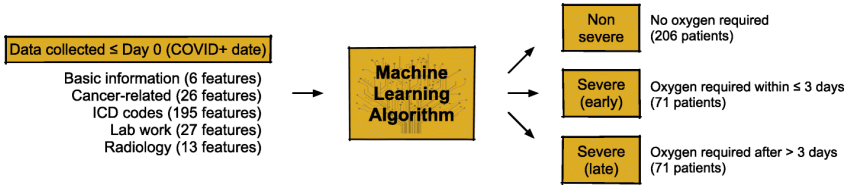
CRISPR-Cas9 negative selection screen (a) and probabilistic model for ACE (b)

Elizabeth R. Hutton, Christopher R. Vakoc, Adam Siepel, "ACE: a probabilistic model for characterizing gene-level essentiality in CRISPR screens" was published *Genome Biology* on September 23, 2021.

A machine learning approach to predicting COVID-19 disease severity

Associate Professor **Saket Navlakha** and his wife, Dr. Sejal Morjaria, an infectious disease physician at Memorial Sloan Kettering Cancer Center (MSKCC), found a way to predict COVID-19 severity in cancer patients. The computational tool they developed prevents unnecessary expensive testing and improves patient care.

The team collected 267 variables from cancer patients diagnosed with COVID-19. The variables ranged from age and sex to cancer type, most recent treatments, and laboratory results. They trained a machine-learning computer program to classify patients into three groups those who will require high oxygen levels through a ventilator either immediately, after a few days, or not at all.



Overview of a machine-learning algorithm for predicting patient outcomes based on data from 348 inpatients at Memorial Sloan Kettering Cancer Center.

The researchers found approximately 50 variables that contributed most to the outcome prediction. Their method had an accuracy rate of 70-95%, and it performed exceptionally well for patients that would require immediate ventilation. More generally, the tool can help tease apart interactions between multiple risk factors that might not be apparent, even to those with trained eyes. The program also prevents over-testing, which will spare patients unnecessary massive hospital costs. Navlakha and Morjaria hope their work will inspire more physicians and computer scientists to work together and create innovative clinical solutions for complex diseases.

Saket Navlakha, Sejal Morjaria, Rocio Perez-Johnston, Allen Zhang, Ying Taur, "Projecting COVID-19 disease severity in cancer patients using purposefully-designed machine learning" was published in *BMC Infectious Diseases* on May 4, 2021.

Assistant Professor Saket Navlakha and Dr. Sejal Morjaria, internist and infectious disease specialist at Memorial Sloan Kettering Cancer Center.

COLLABORATION AND OUTREACH

EVERY YEAR, THE SCQB continues to strengthen its close collaborative ties across CSHL and among NY-area institutions. Eighty percent of our core faculty are now established CSHL Cancer Center members, and many continue to have affiliations with the New York Genome Center (NYGC).

Several SCQB faculty members partnered with traditional CSHL cancer laboratories to explore new innovative quantitative approaches to tumor biology this year. Fellow Hannah Meyer teamed up with Assistant Professor Tobias Janowitz to study cancer immunotherapy. Associate Professor Peter Koo joined forces with Professor Christopher Vakoc to understand sarcoma cancer, and Assistant Professor David McCandlish worked alongside former Fellow Jason Sheltzer to model cancer progression.

Associate Professor **Ivan Iossifov** is still a core faculty member at the NYGC. Professor **Adam Siepel** continues to co-lead the **Population Genomics Working Group** with David Knowles, Ph.D. of the NYGC. This group brings together leading population and statistical geneticists and focuses on human population genomics, statistical analyses, and applications to precision medicine. Faculty members also organized several virtual QB meetings and conferences this year.

Meetings and Conferences

The SCQB faculty members have co-organized the following virtual meetings, conferences, and workshops in 2021.

- 7th **Probabilistic Modeling in Genomics** (**ProbGen '21**), Virtual, April 2021, Co-organized by Adam Siepel
- **FASEB Mobile DNA: Evolution, Diversity and Impact,** Virtual, June 2021, Co-organized by Molly Gale Hammell
- 8th **Biological Distributed Algorithms** (**BDA**), Virtual, July 2021, Co-organized by Saket Navlakha
- **Representation Learning in Biology,** Part of ISMB/ECCB July 2021, Virtual, July 2021, Co-organized by Peter Koo
- **Rita Allen Scholars Symposium,** Virtual, November 2021, Co-organized by Molly Gale Hammell

Interdisciplinary Scholars in Experimental and Quantitative Biology Program

The SCQB introduced the **Interdisciplinary Scholars in Experimental and Quantitative Biology** (ISEQB) program in 2017 as an innovative funding opportunity for postdoctoral research and training. Since then, the ISEQB has funded several talented postdocs interested in both wet-lab and dry-lab research, several of whom have secured NIH funding or landed faculty positions. This interdisciplinary training program promotes collaborative research between experimental and quantitative groups, thereby growing and strengthening the QB community at CSHL.

Saket Navlakha and Hannah Meyer

This year, the SCQB is pleased to award Assistant Professor **Saket Navlakha** and Fellow **Hannah Meyer**, who runs a 50/50 wet-dry lab, funding an ISEQB scholar. The awarded postdoc will explore how T cells distinguish self from non-self in a study entitled, "Compressed sensing in the immune system."

Associate Professor Molly Gale Hammell addresses the Women's Partnership for Science luncheon at the Banbury Center.

THE SCQB SERVES AS A HUB for research, training, and education in the quantitative biological sciences. Students, postdocs, and staff are encouraged to participate in regular symposia, weekly informal gatherings, classes covering advanced quantitative material, and journal clubs focused on cutting-edge research.

Ph.D. training program

Justin Kinney
Lead Instructor

We offer a comprehensive training program in quantitative biology to Ph.D. students through the **Cold Spring Harbor Laboratory School of Biological Sciences**. This program includes a **2.5-day QB Bootcamp** which introduces incoming students to Python programming and high-performance computing, followed by a **22 lecture QB course** taught by a team of QB faculty. The lectures series focuses on a wide range of topics in **machine learning, algorithms, population genetics, functional genomics, image analysis, and biophysics**.

Students who subsequently choose to pursue a Ph.D. in QB are encouraged to participate in the **SCQB Mentored Independent Study** program. Under the guidance of CSHL faculty, students in this program develop and pursue an individualized curriculum of directed reading in quantitative material. Students gain graduate-level training in modern quantitative methods (e.g., Bayesian inference, deep learning, theoretical evolution, dynamical systems, etc.) while learning the skills they will need to continue educating themselves through their scientific careers.

2021 Doctoral Recipients
Cold Spring Harbor School of Biological Sciences

NGOC (TUMI) BUU TRAN, PHD

Dr. Tran conducted her research in Professor Alex Koulakov's laboratory where she built machine learning models to link molecular structures to olfactory perception. Her thesis was entitled, "DeepNose: Using artificial neural networks to represent the space of odorants".

2021 Alumni
Simons Center for Quantitative Biology

WEI-CHIA CHEN, PHD

Dr. Chen, formerly of the Kinney Lab, in now an Assistant Professor in the Department of Physics at National Chung Cheng University in Taiwan.

JUANNAN ZHOU, PHD

Dr. Zhou, formerly of the McCandlish lab, is now an Assistant Professor in the Department of Biology at the University of Florida.

Assistant Professor Peter Koo in his lab with graduate student Shushan Toneyan.

CSHL Fellow Hannah Meyer wins UK Biobank researcher award

Fellow **Hannah Meyer** received the Early Career Researcher of the Year Award for 2021 from the UK Biobank, which honors early-career researchers who have made significant scientific discoveries using their biomedical database. The award recognizes Meyer's study that examined the structure of hearts of over 18,000 individuals in the database. The UK Biobank contains the genetic and medical information from approximately half a million individuals across the United Kingdom. The collected data enables researchers worldwide to understand human health, study disease, and develop treatments.

By combining artificial intelligence and genetic analyses, Meyer and her collaborators discovered the role of cardiac trabeculae, a network of muscle fibers in the adult heart. They showed how these trabeculae allow blood to flow more efficiently and how the muscle pattern can influence the risk of heart failure.

Fellow Hannah Meyer (left) with postdoc Sarah Chapin in her lab.

Assistant Professor Tatiana Engel at the whiteboard in her lab.

NIH Brain Initiative invests $9.7 million in SCQB scientists

The National Institute of Health (NIH) Brain Research through Advancing Innovative Neurotechnologies® (BRAIN) Initiative awarded a total of $9.7 million in new grants to Cold Spring Harbor Laboratory (CSHL) scientists, including Assistant Professor, **Tatiana Engel**. Starting in 2013, the BRAIN Initiative funds innovative neuroscience projects that aim to treat better, prevent, and cure neurological disorders. Engel's laboratory is developing a new theoretical model with former ISEQB Scholar James Roach to understand what happens in the brain when a mouse makes a choice. This computer model will capture the diversity of brain cells needed to enable mice's cognition and action.

APPENDIX

Publications and Preprints from 2021

1. Beyaz, S., Chung, C., Mou, H., Bauer-Rowe, K. E., Xifaras, M. E., Ergin, I., Dohnalova, L., Biton, M., Shekhar, K., Eskiocak, O., Papciak, K., Ozler, K., Almeqdadi, M., Yueh, B., Fein, M., Annamalai, D., Valle-Encinas, E., Erdemir, A., Dogum, K., Alici-Garcipcan, A., **Meyer, HV**,… Yilmaz, Ö. H. (2021). Dietary suppression of MHC class II expression in intestinal epithelial cells enhances intestinal tumorigenesis. *Cell Stem Cell*, 28(11), 1922–1935.e5.

2. Blumberg, A., Zhao, Y., Huang, Y.-F., Dukler, N., Rice, E. J., Chivu, A. G., Krumholz, K., Danko, C. G., & **Siepel, A**. (2021). Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data. *BMC Biology*, 19(1), 30.

3. Cano, A. V., Rozhoňová, H., Stoltzfus, A., **McCandlish, D. M**., & Payne, J. L. (2021). Mutation bias shapes the spectrum of adaptive substitutions. [Preprint] In *bioRxiv* (p. 2021.04.14.438663).

4. Chandrasekhar, A., Marshall, J. A. R., Austin, C., Navlakha, S., & Gordon, D. M. (2021). Better tired than lost: turtle ant trail networks favor coherence over short edges. *PLoS Computational Biology*, 17(10), e1009523.

5. Chen, W.-C., Zhou, J., Sheltzer, J. M., **Kinney, J. B**., & **McCandlish, D. M**. (2021). Field-theoretic density estimation for biological sequence space with applications to 5' splice site diversity and aneuploidy in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 118(40). https://doi.org/10.1073/pnas.2025782118

6. Dewan, R., Chia, R., Ding, J., Hickman, R. A., Stein, T. D., Abramzon, Y., Ahmed, S., Sabir, M. S., Portley, M. K., Tucci, A., Ibáñez, K., Shankaracharya, F. N. U., Keagle, P., Rossi, G., Caroppo, P., Tagliavini, F., Waldo, M. L., Johansson, P. M., Nilsson, C. F., The NYGC ALS Consortium which includes **Gale Hammell, M.**, …Traynor, B. J. (2021). Pathogenic huntingtin repeat expansions in patients with frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron*, *109*(3), 448–460.e4.

7. Dukler, N., Mughal, M. R., Ramani, R., Huang, Y.-F., & **Siepel, A.** (2021). Extreme purifying selection against point mutations in the human genome. [Preprint] In *bioRxiv* (p. 2021.08.23.457339).

8. **Gale Hammell, M**., & Rowe, H. M. (2020). Editorial Overview: Endogenous retroviruses in development and disease. *Viruses*, 12(12).

9. Ghotra, R., Lee, N. K., Tripathy, R., & **Koo, P. K**. (2021). Designing interpretable convolution-based hybrid networks for genomics. [Preprint] In *bioRxiv* (p. 2021.07.13.452181).

10. Ghotra, R. S., Lee, N. K., & **Koo, P. K**. (2021). Uncovering motif interactions from convolutional-attention networks for genomics. [Workshop paper] In the 35th Conference on Neural Information Processing Systems.

11. Hejase, H. A., Mo, Z., Campagna, L., & **Siepel, A.** (2021). A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. [Epub ahead of print] *Molecular Biology and Evolution*, msab332.

12. Henry, S., Trousdell, M. C., Cyrill, S. L., Zhao, Y., Feigman, M. J., Bouhuis, J. M., Aylard, D. A., **Siepel, A.**, & Dos Santos, C. O. (2021). Characterization of gene expression signatures for the identification of cellular heterogeneity in the developing mammary gland. *Journal of Mammary Gland Biology and Neoplasia*, 26(1), 43–66.

13. How, J. J., **Navlakha, S.**, & Chalasani, S. H. (2021). Neural network features distinguish chemosensory stimuli in Caenorhabditis elegans. *PLoS Computational Biology*, 17(11), e1009591.

14. Hutton, E. R., Vakoc, C. R., & **Siepel, A**. (2021). ACE: a probabilistic model for characterizing gene-level essentiality in CRISPR screens. *Genome Biology*, 22(1), 278.

15. Inam, H., Sokirniy, I., Rao, Y., Shah, A., Naeemikia, F., O'Brien, E., Dong, C., **McCandlish, D. M.**, & Pritchard, J. R. (2021). Genomic and experimental evidence that ALKATI does not predict single agent sensitivity to ALK inhibitors. *iScience*, 24(11), 103343.

16. Kawaguchi, R. K., Tang, Z., Fischer, S., Tripathy, R., Koo, P. K., & Gillis, J. (2021). Exploiting marker genes for robust classification and characterization of single-cell chromatin accessibility. [Preprint] In *bioRxiv* (p. 2021.04.01.438068).

17. Kleeman, S. O., Ferrer, M., Demestichas, B., Bankier, S., Lee, H., Heywood, T., Ruusalepp, A., Bjorkegren, J. L. M., Walker, B. R., **Meyer, H. V.**, & Janowitz, T. (2021). Cystatin C is a glucocorticoid response gene predictive of cancer immunotherapy failure. [Preprint] In *bioRxiv*. (p. 2021.08.17.21261668).

18. **Koo, P. K.**, Majdandzic, A., Ploenzke, M., Anand, P., & Paul, S. B. (2021). Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Computational Biology*, 17(5), e1008925.

19. **Koo, P. K.**, & Ploenzke, M. (2021). Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence*, 3(3), 258–266.

20. Labelson, E. L., Tripathy, R., & **Koo, P. K**. (2021). Towards trustworthy explanations with gradient-based attribution methods. [Workshop paper] In the 35th Conference on Neural Information Processing Systems.

21. Lee, N. K., & **Koo, P. K.** (2021). Representation learning of genomic sequence motifs via information maximization. [Workshop paper] In International Conference on Machine Learning.

22. **Levy, D.**, Wang, Z., Moffitt, A. B., & Wigler, M. (2021). Accurate measurement of microsatellite length by disrupting its tandem repeat structure. [Preprint] In bioRxiv (p. 2021.12.09.471828).

23. Li, S., Kendall, J., Park, S., Wang, Z., Alexander, J., Moffitt, A., Ranade, N., Danyko, C., Gegenhuber, B., Fischer, S., Robinson, B. D., Lepor, H., Tollkuhn, J., Gillis, J., Brouzes, E., **Krasnitz, A.**, **Levy, D.**, & Wigler, M. (2020). Copolymerization of single-cell nucleic acids into balls of acrylamide gel. *Genome Research*, 30(1), 49–61.

24. Majdandzic, A., & **Koo, P. K**. (2021). Statistical correction of input gradients for black box models trained with categorical input features. [Workshop paper] In Internation Conference on Machine Learning.

25. **McCandlish, D. M**. (2021). System-specificity of genotype-phenotype map structure: Comment on "From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics" by Susanna Manrubia et al. *Physics of Life Reviews*, 39, 73–75.

26. Mo, Z., Scheben, A., Steinberg, J., **Siepel, A.**, & Martienssen, R. (2021). Circadian immunity, sunrise time and the seasonality of respiratory infections. [Preprint] In *medRxiv* (p. 2021.03.29.21254556).

27. **Navlakha, S**., Morjaria, S., Perez-Johnston, R., Zhang, A., & Taur, Y. (2021). Projecting COVID-19 disease severity in cancer patients using purposefully-designed machine learning. *BMC Infectious Diseases*, 21(1), 391.

28. NeuroLINCS Consortium which includes **Gale Hammell, M**. , Li, J., Lim, R. G., Kaye, J. A., Dardov, V., Coyne, A. N., Wu, J., Milani, P., Cheng, A., Thompson, T. G., Ornelas, L., Frank, A., Adam, M., Banuelos, M. G., Casale, M., Cox, V., Escalante-Chong, R., Daigle, J. G., Gomez, E., … Thompson, L. M. (2021). An integrated multi-omic analysis of iPSC-derived motor neurons from C9ORF72 ALS patients. *iScience*, 24(11), 103221.

29. Petti, S., Bhattacharya, N., Rao, R., Dauparas, J., & Thomas, N., Zhou, J., Rush, AM., **Koo, P.K.**, Ovchinnikov, S. (2021). End-to-end learning of multiple sequence alignments with differentiable Smith-Waterman. [Preprint] In *bioRxiv* (p. 2021.10.23.465204).

30. Scheben, A., Ramos, O. M., Kramer, M., Goodwin, S., Oppenheim, S., Becker, D. J., Schatz, M. C., Simmons, N. B., **Siepel, A.**, & Richard McCombie, W. (2021). Long-read sequencing reveals rapid evolution of immunity- and cancer-related genes in bats. [Preprint] In *bioRxiv* (p. 2020.09.09.290502).

31. Shen, Y., Dasgupta, S., & **Navlakha, S.** (2021). Algorithmic insights on continual learning from fruit flies. [Preprint] In *arXiv* http://arxiv.org/abs/2107.07617

32. Shen, Y., Wang, J., & **Navlakha, S**. (2021). A Correspondence Between Normalization Strategies in Artificial and Biological Neural Networks. *Neural Computation*, 33(12), 3179–3203.

33. **Siepel, A.** (2021). A Unified Probabilistic Modeling Framework for Eukaryotic Transcription Based on Nascent RNA Sequencing Data. [Preprint] In *bioRxiv* (p. 2021.01.12.426408).

34. Skalenko, K. S., Li, L., Zhang, Y., Vvedenskaya, I. O., Winkelman, J. T., Cope, A. L., Taylor, D. M., Shah, P., Ebright, R. H., **Kinney, J. B.**, Zhang, Y., & Nickels, B. E. (2021). Promoter-sequence determinants and structural basis of primer-dependent transcription initiation in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 118(27).

35. Williams, E. H., Flint, T. R., Connell, C. M., Giglio, D., Lee, H., Ha, T., Gablenz, E., Bird, N. J., Weaver, J. M. J., Potts, H., Whitley, C. T., Bookman, M. A., Lynch, A. G., **Meyer, H. V.**, Tavaré, S., & Janowitz, T. (2021). CamGFR v2: A new model for estimating the glomerular filtration rate from standardized or non-standardized creatinine in patients with cancer. *Clinical Cancer Research*. 27(5), 1381 1390.

36. Wu, Y., Johnson, L., Song, B., Romay, C., Stitzer, M., **Siepel, A.**, Buckler, E., & Scheben, A. (2021). A multiple genome alignment workflow shows the impact of repeat masking and parameter tuning on alignment of functional regions in plants. [Preprint] In *bioRxiv* (p. 2021.06.01.446647).

37. Yoon, S., Munoz, A., Yamrom, B., Lee, Y.-H., Andrews, P., Marks, S., Wang, Z., Reeves, C., Winterkorn, L., Krieger, A. M., Buja, A., Pradhan, K., Ronemus, M., Baldwin, K. K., **Levy, D.**, Wigler, M., & Iossifov, I. (2021). Rates of contributory de novo mutation in high and low-risk autism families. *Communications Biology*, 4(1), 1026.

38. Liang, Y., Ryali, CK, Hoover, B., Grinberg,L., **Navlakha, S.**, Zaki, MJ., Krotov. D. (2021). Can a Fruit Fly Learn Word Embeddings? [Preprint] *arXiv* https://arxiv.org/abs/2101.06887

39. Zhao, Y., Dukler, N., Barshad, G., Toneyan, S., Danko, C. G., & **Siepel, A.** (2021). Deconvolution of Expression for Nascent RNA sequencing data (DENR) highlights pre-RNA isoform diversity in human cells. [Epub ahead of print] *Bioinformatics*, btab582.

40. Ziamtsov, I., Faizi, K., & **Navlakha, S.** (2021). Branch-pipe: Improving graph skeletonization around branch points in 3D point clouds. Remote Sensing, 13(19), 3802.