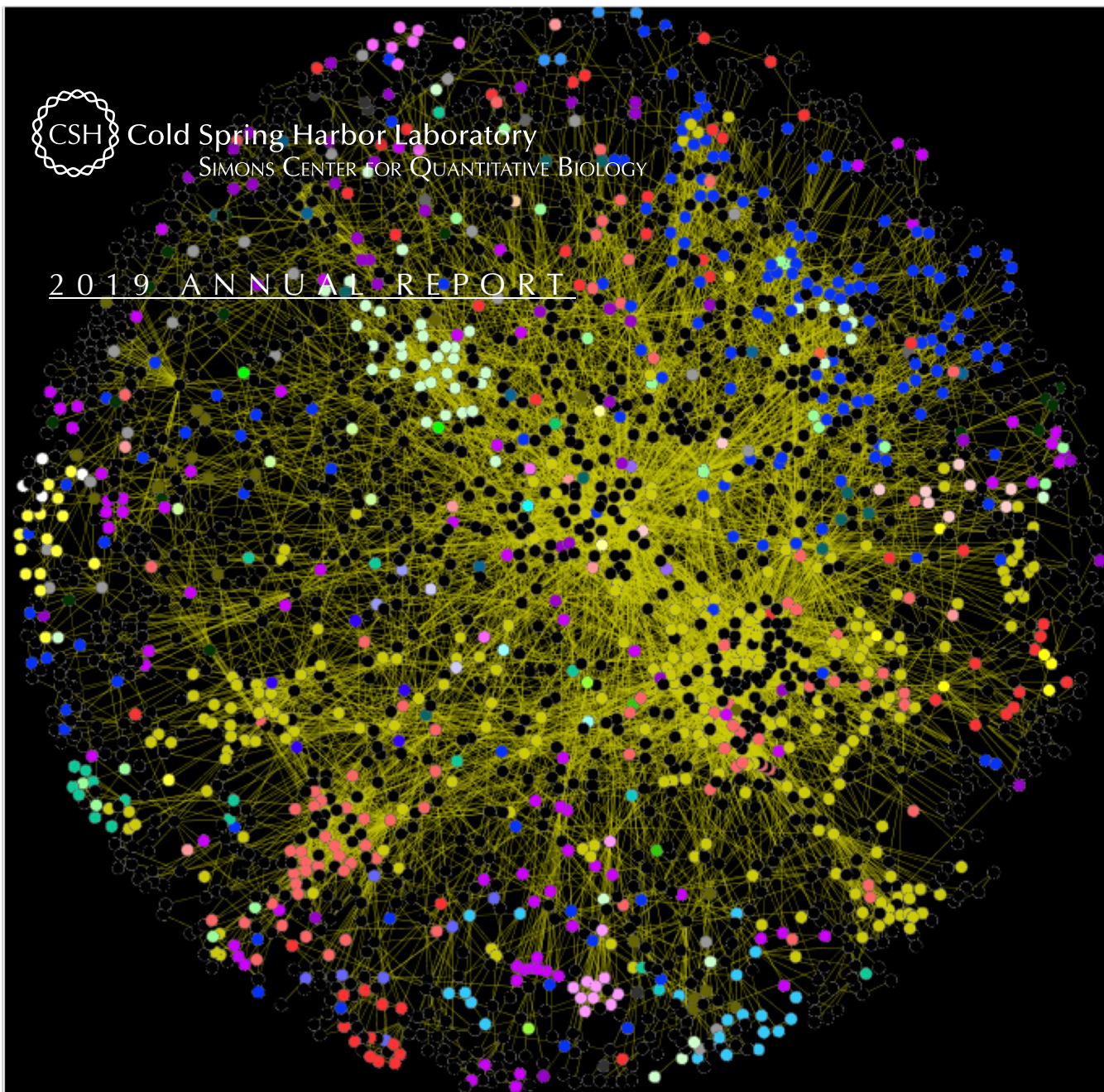




Cold Spring Harbor Laboratory
SIMONS CENTER FOR QUANTITATIVE BIOLOGY

2019 ANNUAL REPORT



Cover image: Protein interaction network for the yeast *S. cerevisiae*, credit Saket Navlakha.

LETTER FROM THE CHAIR

LAST MONTH, WE HELD OUR ANNUAL CSHL In-House Symposium, where faculty members from across the Lab present their recent research activities. As always, it was a stimulating and



Adam Siepel

rewarding event, with reports of exciting new findings across diverse scientific areas. What really struck me this year, however, was the degree to which quantitative biology has permeated research at CSHL. The Symposium included several excellent talks from both new and established members of our QB faculty. Even more, it was clear from many other talks that the use of tools such as machine learning, data science, and mathematical modeling have become central to research across the Laboratory, spanning areas such as Plant Biology, Neuroscience, and Cancer Biology. I took this emerging prominence of quantitative biology as a sign that we are doing something right in our efforts to grow and strengthen a community at CSHL that is focused on the development and application of quantitative methods to problems in biological research.

This past year, in particular, has brought with it a number of important new developments and accomplishments at the SCQB. Perhaps most important, we were very pleased after several years of intensive recruiting to have three outstanding new investigators arrive at the center: Associate Professor Saket Navlakha, Assistant Professor Peter Koo, and CSHL Fellow Hannah Meyer. As detailed on p. 8, these three individuals strengthen the research portfolio of the center in critically important areas, complement our existing faculty, and to add to its diversity. In addition, one of our junior faculty members, Justin Kinney, was promoted this year to Associate Professor, with strong support from the senior faculty at the Laboratory. His promotion brings our number of Associate Professors to six (see p. 3), and highlights that our faculty which was disproportionately junior when I arrived five years ago is steadily becoming more established.

In another sign that the group is reaching its stride, both Justin Kinney and David McCandlish were successful in obtaining major five-year Maximizing Investigators' Research Award (MIRA) grants from the NIH this year, and Hannah Meyer won her first grant, an award from the Mathers Foundation (p. 8). Additional grants were awarded to Alex Krasnitz, Ivan Iossifov, Saket Navlakha, and myself, for a total of \$8.8M in new awards. In

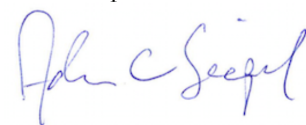
addition, the group produced 41 new publications and preprints, spanning all of our major thematic areas: gene regulation, evolutionary genomics, genomic disease research, and genomic technology development (p. 3; see highlighted projects pp. 4–5).

Another major focus this year has been on improving collaborative ties both within CSHL and with other institutions. Toward this end, we have strengthened our ties to the New York Genome Center through the involvement of Ivan Iossifov, Molly Hammell, and myself in activities there (p. 6). In addition, I am working closely with Prof. David Tuveson, Director of the CSHL Cancer Center, to better integrate quantitative biology in cancer research at CSHL. In particular, we are in the process of adding more QB faculty members to the roster of the Cancer Center, and of offering seed funding for collaborative projects between quantitative biologists and experimentalists. Prof. Tuveson and I believe there are many rich opportunities for collaboration that can benefit both our quantitative and experimental research programs, and substantially advance cancer research at CSHL.

Beyond research, we have maintained a strong commitment to training graduate students and postdoctoral associates in quantitative biology at the SCQB (p. 7). We offered two advanced courses this year—an online course in Machine Learning led by SCQB postdocs Ammar Tareen and Noah Dukler, and a reading group on Gaussian Processes led by David McCandlish and SCQB affiliate Tatiana Engel. In addition, Justin Kinney, as the faculty lead for the QB Course for the Watson School, has begun a major reorganization of our curriculum and recruited several new faculty members to participate in teaching this important introductory course. Finally, the center continues to produce new PhD recipients at a good clip, with Drs. Melissa Hubisz and Kristina Grigaityte graduating in 2019 (p. 7).

Altogether, 2019 has been an exciting and productive year, and we look forward to continuing success in 2020.

Best Wishes for the New Year,
Adam Siepel, PhD, Chair

A handwritten signature in blue ink that reads "Adam Siepel".

December 31, 2019

RESEARCH ACTIVITIES

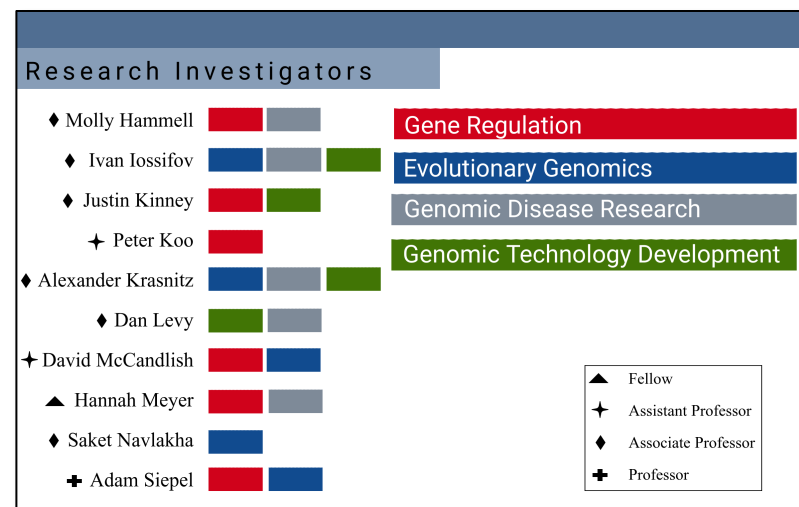
THE SIMONS CENTER FOR QUANTITATIVE BIOLOGY IS focused broadly on revealing how genomes work, how they evolve, and what makes them go wrong in disease. Investigators at the SCQB pursue diverse research interests in a wide variety of different areas, but our research continues to be permeated by four major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

GENE REGULATION

Several researchers at the SCQB are interested in developing both theoretical and experimental methods, along with computational and mathematical 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 behavior of small non-coding RNAs, inference of gene regulatory networks, 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

Other scientists at the SCQB develop theory and mathematics to address a number of open questions in evolutionary genetics, including the dynamics of evolution when mutation is rate-limiting or exhibits biased patterns, and the evolutionary implications of epistasis, i.e. interactions between mutations and genes. Additional studies at the Center use evolutionary methods to identify regulatory elements, to reconstruct early human history, and to estimate the fitness consequences of new mutations in the human genome. Researchers also use evolutionary signatures to aid in the identification of genes associated with autism spectrum disorder and employ phylogenetic methods to study the evolution of tumors.



GENOMIC DISEASE RESEARCH

Several researchers at the Center are trying to understand the genetics of autism spectrum disorder (ASD) through the analysis of large genomic data sets while other researchers are developing mathematical and statistical tools to characterize the cellular composition, genomic disruptions, evolutionary history, and invasive capacity of malignant tumors. In addition, scientists at the Center are investigating the role of transposable element activation in neurodegenerative diseases, particularly amyotrophic lateral sclerosis (ALS) and fronto-temporal dementia (FTD). Researchers are also interested in diverse modeling and statistical inference problems having to do with cancer and immunology, often through consideration of single-cell sequencing data. Several members of the SCQB are active participants in the CSHL Cancer Center.

GENOMIC TECHNOLOGY DEVELOPMENT

Various research groups in the SCQB are working on the development of 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.

RESEARCH HIGHLIGHTS

IN 2019 MEMBERS OF THE SCQB, published research articles in a diverse collection of leading scientific journals. The following is a sampling of this year's important findings.

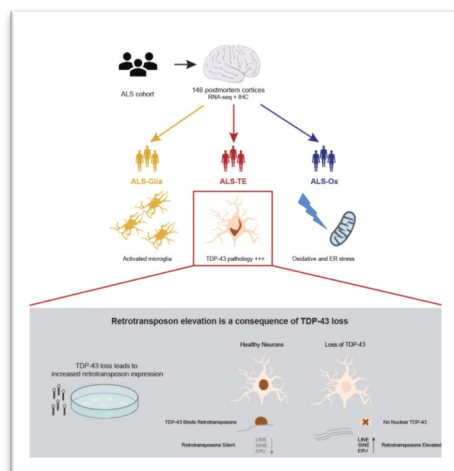
Jumping genes found in ALS suggest possible therapeutic directions



Molly Gale Hammell

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disorder where few patients carry identifiable heritable mutations or a family history of disease. ALS eventually causes paralysis as a protein called TDP-43 collects in nerve cells of the brain and spinal cord, causing these cells to die off. There is no known cure but new research could lead to better diagnosis and more effective treatments. Molly Gale Hammell led a study in collaboration with other leading institutions, including the New York Genome Center,

where researchers aimed to identify different types of ALS patients using a large cohort of ALS patient samples. A large subset of the ALS patients showed strong links between the function of TDP-43 and activation of retrotransposons, also known as “jumping genes”. These jumping genes can randomly move from one spot on the chromosome to another, altering gene expression and thereby influencing the functional activities of the cell. TDP-43 is one of the proteins that keep jumping genes silent but when it accumulates in the nerve cells of ALS patients, TDP-43 fails to silence jumping genes. The team applied machine learning algorithms to study



The different subtypes of ALS and the connection between TDP-43 pathology and jumping genes.

gene expression patterns in the brain tissues of post-mortem patients and saw there was a de-silencing of jumping genes for the subset of patients with the most extensive TDP-43 pathology. Hammell speculates that high levels of jumping genes may be mimicking a viral pathogen or other infection. The team now wants to confirm whether jumping genes are contributing to cell death in ALS patients, which would allow researchers to directly target those genes with antiviral agents or other therapies.

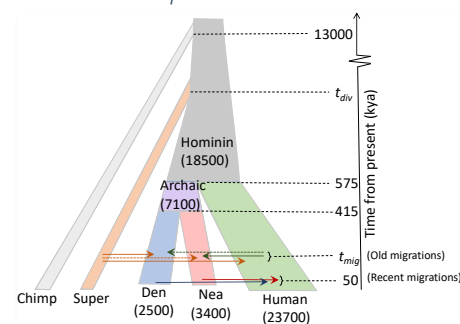
Oliver H. Tam, Nikolay V. Rozhkov, Regina Shaw, Duyang Kim, Isabel Hubbard, Samantha Fennessey ... Molly Gale Hammell, “Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia” was published in *Cell Reports* on October 29, 2019.

Reconstruction of human evolution from genetic data



Adam Siepel

Modern humans harbor genetic fragments from other closely related but long-extinct lineages revealing a history characterized by far more diversity, movement, and mixture than seemed imaginable a mere decade ago. Adam Siepel and his colleagues were particularly interested in looking for signs of gene flow from modern humans into



Population model used for ARGweaver-D analysis.

Neanderthals. That flow of genetic information is harder to study than the reverse, not only because of how long ago it happened but also because there are only a handful of Neanderthal genomes to refer to, prompting the need for new methods. Siepel and his team developed one such new technique, called ARGweaver-D, and used it to show that around 3% of Neanderthal DNA and possibly as much as 6% came from modern humans who mated with Neanderthals more than 200,000 years ago. Therefore, while human/Neanderthal interbreeding contributed to the DNA of modern humans throughout the world, it also resulted in the transfer of genetic variation back to Neanderthals. According to Siepel's analysis, a similar type of nested mixing also happened with the Denisovans, another group of archaic hominins. When the team looked at the Denisovan genome, they found fragments of DNA from an even earlier hominin, vestiges of some population whose own genome has not been found or sequenced. It might have been *Homo erectus*, which split off from ancestors of modern humans and spread across Eurasia roughly 1 million years ago. According to Siepel, the contribution from this unidentified group was at the limits of their detection power and constituted only about 1% of the Denisovan genome. During later interbreeding events, tiny pieces of that 1% got passed on to modern humans in Southeast Asia, Papua New Guinea, and some parts of East Asia. If the analysis done by Siepel's team is correct, some extremely

divergent DNA sequences present in modern humans would have been passed through two interbreeding events involving archaic hominins.

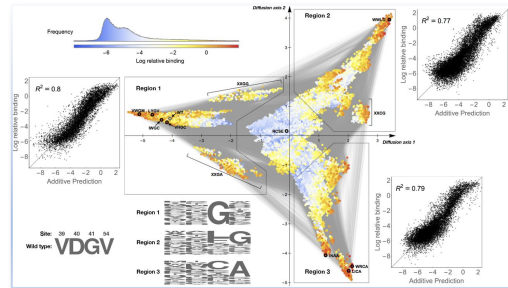
Melissa J. Hubisz, Amy L. Williams and Adam Siepel, “Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph” was posted to *bioRxiv* June 30, 2019.

Modeling how multiple mutations combine to impact biological function



David McCandlish

Massively parallel phenotyping assays use high-throughput DNA sequencing to measure the molecular phenotypes of a large number of sequences, providing unprecedented insight into how mutations combine to determine biological function. Due to the vastness of sequence space and limitations in scalability, these assays typically result in missing sequences for many possible genotypes. Making accurate phenotypic predictions for these missing sequences is a difficult problem because the effect of any given mutation often depends on which other mutations are already present in the sequence, a phenomenon known as epistasis. Assistant Professor David McCandlish set out to develop a technique that can accurately predict values for genotypes whose phenotypes are not directly assayed. In a recent study,



Visualization of the GBI landscape reconstructed using minimum epistasis interpolation and the local non-epistatic smoother.

McCandlish and postdoc Juannan Zhou developed a method that infers a set of predictions in which mutational effects change as little as possible across adjacent genetic backgrounds, resulting in improved prediction power. They applied this method to analyze a “fitness landscape” (genotype-fitness map) for the IgG binding domain of streptococcal protein G, a model system for studying protein folding stability and binding affinity.

McCandlish and Zhou were able to show that their technique can explain this landscape with less epistasis but comparable predictive power to standard methods. Moreover, their analysis reveals that the complex structure of epistasis observed in this dataset can be well-understood in terms of a simple qualitative model.

Juannan Zhou and David McCandlish, “Minimum epistasis interpolation for sequence-function relationships” was posted to *bioRxiv* June 4, 2019.

Genomic analysis reveals properties of heart muscle crucial to cardiovascular function

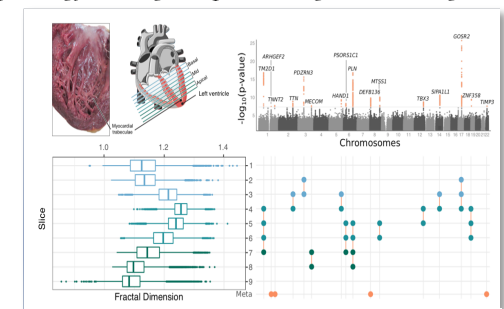


Hannah Meyer

The inner surfaces of the adult heart are covered by a complex network of muscular strands. These strands, known as muscular trabeculae, are thought to be a vestige of embryonic development. Their function in adults and their genetic architecture are unknown. In a recent study, led by CSHL Fellow Hannah Meyer, researchers harnessed population genomics, image-based phenotyping and *in silico* modeling to understand the effect of this image-derived trait on cardiovascular function and physiology. Using deep learning-based image analysis, Meyer identified genetic regions associations with trabecular complexity in 18,097 UK Biobank participants which were replicated in two independently measured cohorts. Genes in these regions are significantly enriched for expression in the fetal heart or vasculature. These regions are also associated with phenotypes related to the flow of blood within tissues and related developmental pathways. Meyer

found a causal relationship between the increasing complexity of these muscular networks and both ventricular performance and cardiovascular conductivity. This relationship is supported by complementary biomechanical simulations, Mendelian randomization studies and medaka fish knock-out models. Overall, Meyer’s findings demonstrate that trabeculae complexity is a powerful determinant of cardiac performance, and that associated genetic variants affect susceptibility to heart failure and hypertension. These findings show how genetically-determined structural complexity in the adult heart is an adaptation crucial to cardiovascular function.

Hannah V. Meyer et. al., “Genomic Analysis reveals a functional role for myocardial trabeculae in adults” was posted to *bioRxiv* March 1, 2019.



Complexity (lower left) of muscular network on inner surface of the heart muscle (upper left) is associated with 16 independent loci across the genome (upper right) which are specific for distinct regions in the ventricle (lower right).

COLLABORATIONS

MEMBERS OF THE SCQB maintain close collaborative ties across CSHL and with many other New York area groups. Faculty members also organize relevant QB meetings and conferences at CSHL and around the NY area.

OTHER AFFILIATIONS AND ORGANIZATIONS

The New York Genome Center



As one of the New York Genome Center's (NYGC) founding institutional members, Cold Spring Harbor Laboratory (CSHL) share interests in a variety of scientific topic areas including, autism spectrum disorder, neurodegenerative disease, and cancer.

Several SCQB faculty members maintain close affiliations with the NYGC.



Ivan Iossifov As a core NYGC member since 2015, Assistant Professor **Ivan Iossifov** oversees several large autism screening projects and splits his time between the NYGC and CSHL. Associate Professor **Molly Hammell** is a member of the NYGC ALS Consortium, a worldwide collaborative effort initiated to advance study of the disease. Dr. Hammell's work with the NYGC ALS Consortium resulted in a recent publication in *Cell Reports* as previously detailed in this report.

Professor **Adam Siepel** continues to co-lead the **Populations Genomics Working Group** at the NYGC with Eimear Kenny, PhD, of the Icahn School of Medicine at Mount Sinai. Among other activities, the Populations Genomics Working group features a series of talks on evolutionary and population genomics, computational biology, and machine learning. One of these talks was recently delivered by Assistant Professor **Peter Koo** who presented his work on "Interpretable Deep Learning for Regulatory Genomics" in November 2019.

Meetings and Conferences

The SCQB faculty members have co-organized the following meetings and conferences in 2019.

- **New York Area Population Genomics Workshop**, Icahn School of Medicine at Mount Sinai, New York, NY, January 2019, Co-organized by Adam Siepel
- **Biological Distributed Algorithms Workshop**, DoubleTree Hilton Hotel Toronto Downtown, Toronto, Canada, July 2019, co-chaired by Saket Navlakha

Work is underway to organize several events in the future including the **Rita Allen Foundation Scholars Symposium** (Co-organized by Molly Gale Hammell), **FASEB Mobile Genetics** (Co-organized by Molly Gale Hammell), and the **NY Area Quantitative Biology Meeting** (Co-organized by Peter Koo and Saket Navlakha).

Interdisciplinary Scholars in Experimental and Quantitative Biology Program (ISEQB)

The Interdisciplinary Scholars Program in Experimental and Quantitative Biology (ISEQB) is an innovative funding opportunity for postdoctoral research open to applications in all areas of research at CSHL, including genetics, cancer plant biology, and neuroscience. The ISEQB is designed to help recruit new postdocs or fund existing CSHL postdocs who are interested in both wet-lab and dry-lab research. This program aims to catalyze collaborative research as well as promote growth of the QB community at CSHL.



Walter Bast
Postdoctoral Researcher

Neuroscience - Odor Perception

Supervised by:
Alexei Koulakov and Florain Albeanu



James Roach
Postdoctoral Researcher

Neuroscience - Decision Making

Supervised by:
Tatiana Engel and Anne Churchland



Wei Chia Chen
Postdoctoral Researcher

Clinical Data Analysis

Supervised by:
Justin Kinney and Robert Maki



Mandy Wong
Postdoctoral Researcher

RNA Splicing

Supervised by:
Justin Kinney and Adrian Krainer

EDUCATION AND OUTREACH

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.

PhD training program

We continue to offer a broad training program in quantitative biology to PhD students through the Watson School of Biological Sciences (WSBS).

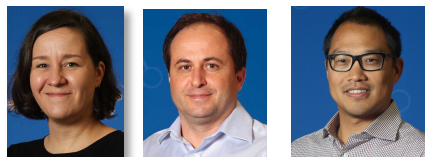


Justin Kinney, QB curriculum director

Incoming students first take a **2.5-day QB Bootcamp** that introduces them to Python programming and high-performance computing. This is followed by a **22 lecture QB course** which is taught by a team of QB faculty. The first part of this course focuses on essential biomedical statistics as well as on Python programming. This is followed by introductory lectures on a wide range of topics in machine learning, algorithms, population genetics, functional genomics, image analysis and biophysics.



Beginning in 2019, WSBS students who choose to pursue their research in the SCQB are encouraged to participate in a mentored independent study program. In this program, students work individually or in small groups with a supervising faculty member to design a program of directed reading in quantitative graduate-level course material. Students then pursue these studies in parallel with their thesis research throughout the remainder of their time in the WSBS.

QB course instructors include: Associate Professors Justin Kinney (lead instructor) and **Molly Gale Hammell**, Professor Adam Siepel, Assistant Professors **Alex Dobin**, **David McCandlish**, **Peter Koo**, and Research Assistant Professor Jon Preall.

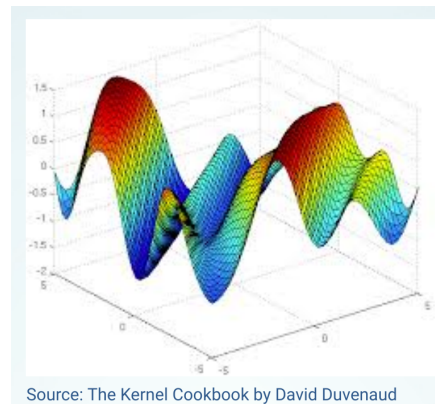


Molly Gale Hammell, Alex Dobin, Peter Koo

2019 DOCTORAL RECIPIENTS

	Advisor	Academic Affiliation	Thesis
 Melissa Hubisz	Adam Siepel	Cornell University	Inferring the population history of ancient hominins through use of the ancestral recombination graph
 Kristina Grigaityte	Mickey Atwal	Watson School of Biological Sciences	Comprehensive sequencing analyses of high-throughput single T cells in humans

Advanced QB Course: Gaussian Processes for Biological Data Analysis



David McCandlish and Tatiana Engel

This year, Assistant Professors, David McCandlish and Tatiana Engel ran an **Advanced Quantitative Biology Course in Gaussian Processes**. This course, open to all members of the CSHL community, explored applications of these techniques to topics of current interest at CSHL, including statistical analysis of neural recordings, large-scale mutagenesis experiments, and clinical trials.

AWARDS AND RECOGNITION

David McCandlish named Sloan Research Fellow



Assistant Professor **David McCandlish** was recently named a 2019 Sloan Research Fellow. Dr. McCandlish develops computational and mathematical tools to analyze genetic data. His lab focuses specifically on analyzing data from so-called “deep mutational scanning” experiments,

which determine, for a single protein, the functional effects of thousands of mutations. Awarded by the Alfred P. Sloan Foundation, Sloan Research Fellowships recognize early-career scholars as exceptionally promising researchers in their fields.

Hannah Meyer receives grant from the Mathers Foundation



Hannah Meyer and David McCandlish
(Photo courtesy: Getty Images)

CSHL Fellow **Hannah Meyer** received a grant from the Mathers Foundation to study how the organization of the thymus and the cells therein provide effective education for T cells through genomics and mathematical modeling. Specifically, Dr. Meyer will develop methods to analyze 3D spatial transcriptomics data and embed these in a larger framework to study cell-cell interactions during T cell development in the thymus.



Justin Kinney and David McCandlish awarded NIH Funding for Early Stage Investigators



Associate Professor **Justin Kinney** and Assistant Professor **David McCandlish** each received an NIH Maximizing Investigators' Research Award (MIRA) for Early Stage Investigators. MIRA grants are designed to provide investigators with greater stability and flexibility than standard NIH grants, thereby enhancing scientific productivity and the chances for important breakthroughs. Dr. Kinney's

award focuses on biophysical modeling of cis-regulatory complexes in transcription and splicing using massively parallel reporter assays. Dr. McCandlish's award focuses on understanding complex genetic interactions through computational analysis.

Molly Gale Hammell awarded 2019 WiSE faculty mentor award

Associate Professor **Molly Gale Hammell** was awarded the 2019 CSHL Women in Science and Engineering (WiSE) Mentorship Award. WiSE was founded in 2015 by students, postdocs, and technicians looking to create a strong and collaborative support system for women scientists at CSHL and beyond. Dr. Hammell received multiple nominations, and her award was presented by one of her nominators, a Watson School of Biological Sciences (WSBS) graduate student in her lab, Kat O'Neill.



Molly Gale Hammell and Kat O'Neill

The SCQB Welcomes Three New Faculty Members



Peter Koo

Dr. Peter Koo joined the SCQB as an Assistant Professor in mid-September 2019 to study the functional impact of genomic variants using data-driven artificial intelligence (AI) solutions. Koo is broadly interested in applications for studying gene regulation and protein (dys)function.



Hannah Meyer

Dr. Hannah Meyer joined the SCQB as a Quantitative Biology Fellow in March 2019. Meyer is interested in how T cells are educated within the thymus gland during development to ultimately react against foreign pathogens. By combining genomics and mathematical modeling, Meyer investigates gene expression in the thymus and the effects of cell-cell interaction during T cell development to understand autoimmunity and autoimmune disease.



Saket Navlakha

Dr. Saket Navlakha moved from the Salk Institute to CSHL, joining the SCQB as an Associate Professor in November 2019. Navlakha is broadly interested in the strategies biological systems have evolved to solve problems that inhibit their survival. By viewing these strategies as algorithms, Navlakha studies neural circuit computation and plant architecture optimization.