2022 ANNUAL REPORT

OF THE SIMONS CENTER FOR QUANTITATIVE BIOLOGY AT COLD SPRING HARBOR LABORATORY

Cold Spring Harbor Laboratory
Simons Center for Quantitative Biology

Cover Image: A neural network in the shape of a gopher was generated by graduate students Shushan Toneyan and Ziqi (Amber) Tang using DALL-E2, a new AI system that can create realistic images and art from a description in natural language.

2022 SCQB ANNUAL REPORT

ADVISORY BOARD

TABLE OF CONTENTS

CORE FACULTY

Molly Gale Hammell
Ivan Iossifov
Justin Kinney
Peter Koo
Alexander Krasnitz
Dan Levy
David McCandlish
Hannah Meyer
Saket Navlakha
Adam Siepel

From the Chair

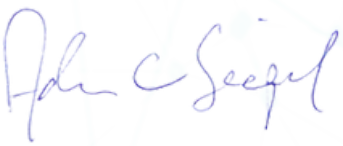
It is astonishing how much scientific life has changed in just a year. In December of 2021, we were experiencing a dispiriting return to near-lockdown conditions with the Omicron surge in COVID-19. We had resumed mask-wearing and limitations on face-to-face meetings. Our long-anticipated return to scientific travel and meeting attendance had been postponed. Now, one year later, the daily rhythm of our activity is much closer to what it was in the pre-pandemic area. Almost everyone is back on campus; face-to-face meetings and teaching have resumed, mostly without masks; the CSHL meetings and courses program is in full swing; and travel to scientific meetings and conferences is rapidly resuming. This return to relative normality has been a breath of fresh air for everyone.

One unexpected benefit of the pandemic is that it provided uninterrupted time for grant writing. Consequently, we had our best year for grants in 2022, with five new awards worth nearly $10M (details on pp. 17 and 22). In recent years, we have been especially effective in obtaining grants from the US National Institutes of Health and National Science Foundation, and in 2022 reached a total of seven active federal grants (p. 17). Our rate of publication slowed a bit from its height during the pandemic but remained strong, with 32 papers and preprints in 2022 (pp. 21–23). In this report, we highlight five recently published studies covering a wide range of research areas, from evolution to cancer to neuroscience (pp. 6–10).

Despite the challenges of the pandemic, we have remained highly active in training students and postdoctoral associates in quantitative biology. We continue to offer extensive training opportunities, including formal courses and more informal mentoring and support (p. 11–12). Three new Ph.Ds were awarded this year to QB students (p. 11). Our students are not only excelling academically but also frequently serving as lead authors on cutting-edge research publications (p. 13). The number of postdoctoral associates in QB labs has also grown over time, and five postdocs successfully moved on to new positions this year (p. 14). We recently funded a new postdoctoral scholar position through our Interdisciplinary Scholars in Experimental Biology (ISEQB) program (pp. 14 and 25), which has previously funded four other highly successful candidates.

It has been a great pleasure to resume face-to-face meetings with our colleagues at other institutions. Last October, we finally held a long-planned one-day symposium with our Simons-funded colleagues at the Flatiron Institute in Manhattan in what proved to be a stimulating and productive exchange of ideas (p. 15). In-person events have also resumed at the NY Genome Center and at other meetings (p. 15). In the coming months, we look forward to helping to organize the New York Area Population Genetics workshop, to be held at Rockefeller University in January, and the Probabilistic Modeling in Genomics (ProbGen) meeting, to be held at CSHL in March.

Best Wishes for the New Year,



Adam Siepel. Ph.D., Chair
December 23, 2022

About the Center

The **Simons Center for Quantitative Biology (SCQB)** was created in 2008 to be the home for mathematical, computational, and theoretical research in biology at Cold Spring Harbor Laboratory (CSHL). As one of the world's leading biomedical research institutions, CSHL has been a leader in recognizing the importance of quantitation in the life sciences.

Chair of CSHL's Board of Trustees Marilyn Simons, Fellow Hannah Meyer, and CSHL President Bruce Stillman.

Quantitative methods in genetics, biophysics, neuroscience, and other areas have always been an important part of the research at the Laboratory.

Our ten core faculty are experts in applied mathematics, computer science, theoretical physics, and engineering who address open problems in biology, often in close collaboration with experimentalists, to further basic research and investigation into illnesses including cancer, autism, bipolar disorder, and neurodegenerative disease.

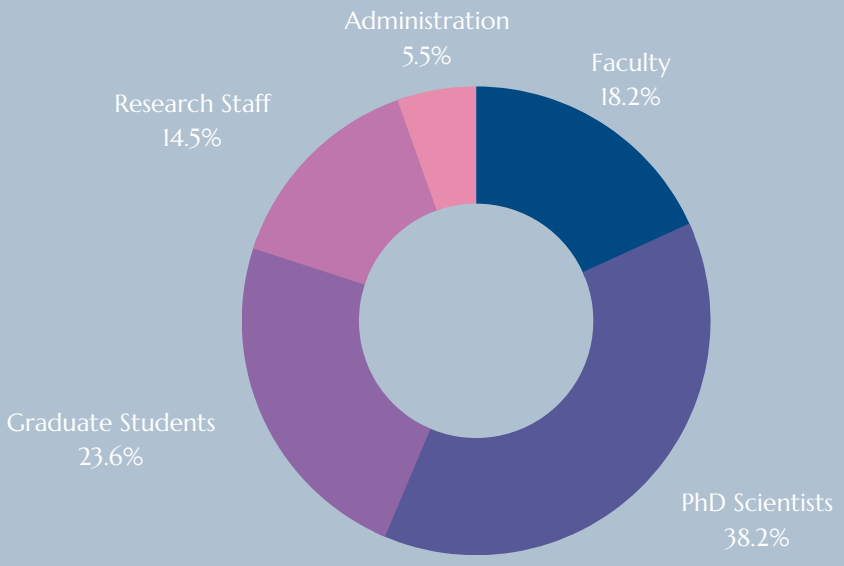
A generous endowment from the Simons Foundation supports the SCQB.

Cold Spring Harbor Laboratory
Simons Center for Quantitative Biology

WHO WE ARE

55

Total Members



Salomé Carcy (left) is a Ph.D. student in Fellow **Hannah Meyer**'s lab who studies T cell development in the thymus through single-cell transcriptomics and multi-omics approaches.

Our Research Themes

1 Gene Regulation

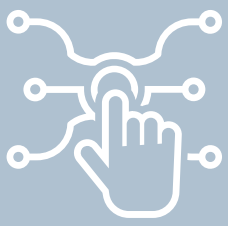
Untangling the relationship between biological sequence and biological function

2 Evolutionary Genomics

Understanding the evolution of gene expression and the evolutionary implications of mutations.

3 Genomic Disease Research

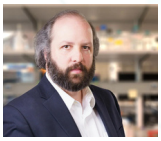
Gaining insights into the genetics of disease through "big data".



4 Genomic Technology Development

Developing new methods, technologies, and systems to analyze genomic information.

CSH Cold Spring Harbor Laboratory
Simons Center for Quantitative Biology

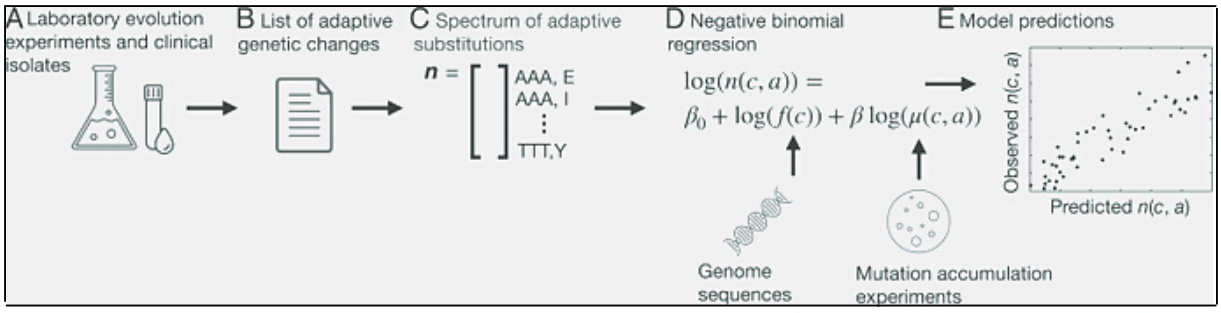


Getting a step ahead of bacteria's drug resistance evolution

In the face of natural selection, genetic mutations set some individuals up for success. A single change to the sequence of a gene can give an organism an advantage, and passing on that mutation to future generations may shape the fate of the species. Scientists have now determined that an imbalance in the types of mutations that often crop up in genomes partly steers evolutionary adaptation. For various reasons, specific segments of the genetic code are more prone to mutation than others. Understanding the impact of this mutational bias could help scientists predict how pathogens and cancer cells are most likely to evolve resistance to the drugs used on them. That, in turn, will aid the design of new medicines and treatment strategies. **Assistant Professor David McCandlish**, who led a study with collaborator Joshua Payne at ETH Zürich, points out that there is often more than one genetic solution to a problem. For example, in bacteria exposed to an antibiotic, dozens of different mutations can confer an ability to resist the drug, but certain paths to resistance are more common than others.

Hundreds of different drug-resistance mutations have been found in tuberculosis (TB) bacteria isolated from patients, but some of these are 100 times more common than others. McCandlish and his colleagues analyzed thousands of adaptive changes found in three different microorganisms, including bacteria that cause tuberculosis. The researchers determined that the types of genetic changes accumulated during adaptive evolution, both in natural and laboratory settings, are the mutations that occur the most frequently. They reported these findings in the **Proceedings of the National Academy of Sciences**, noting several factors that influence how closely the adaptive changes found in a population mirror the spectrum of mutations in that species. The researchers discovered that the relationship was strongest when populations experienced few beneficial mutations per generation. In other words, if the number of mutations is low, the population is just waiting for any mutation that can get the job done. If the number is high, there are likely to be several different mutations that can solve this problem, and the one mutation we discover at the end of the process is the one that solves it best.

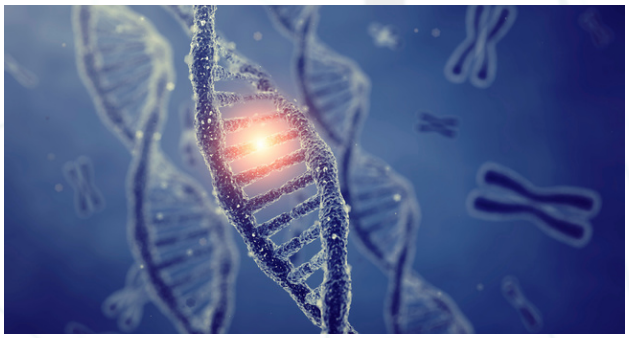
Alejandro V. Cano, Hana Rozhoňová, Arlin Stoltzfus, David M. McCandlish & Joshua L. Payne, "Mutation bias shapes the spectrum of adaptative substitutions" was published in *Proceedings of the National Academy of Sciences (PNAS)* on February 15, 2022.



Workflow for predicting the spectrum of beneficial mutations.

Exposing the evolutionary weak spots in the human genome

Mutations can drastically help or hurt the odds of an organism surviving and reproducing. **Professor Adam Siepel's** team has created a computational method called ExtRaINSIGHT, that tracks the history of deleterious mutations in the human genome throughout evolution. Siepel's method combines deep sequencing data and classical theory from population genetics to discover harmful mutations by looking for their absence. By random chance, every part of the human genome should have mutations, but some have none. Siepel calls these places "ultraselected." When they occur, the mutations there can be deadly or drastically hurt the odds of reproducing. If one were to look at a panel of a hundred thousand humans and never see a mutation at a particular gene, this would suggest that any mutation that did occur was so harmful anyone carrying that mutation died out from the population. The team analyzed over 70,000 human genomes with ExtRaINSIGHT and discovered three parts of the human genome have been extremely sensitive to mutations over generations. These include splice sites, ancient miRNAs, and neuronal protein-coding genes. Splice sites are the most sensitive and help produce correct protein-making instructions. Mutations in a splice site can significantly impact the odds of passing down genes, known as fitness, and are linked to several diseases, including spinal muscular atrophy, the leading genetic cause of death in infants and toddlers.



Splice site mutations can reduce fitness by 1 or 2%, and the accumulation of multiple splice site mutations eventually limits the chances of passing on genes to zero. ExtRaINSIGHT revealed that miRNAs are also sensitive, which made sense to Siepel and his team since miRNAs are often important in critical developmental pathways.

Many genetic diseases and conditions remain a mystery, and Siepel hopes technologies like ExtRaINSIGHT will help reveal their origins and guide diagnoses and treatments.

The researchers were also unsurprised to learn that neuronal-protein coding genes are particularly sensitive to mutation, given the complexity and interconnectedness of the nervous system. Many genetic diseases and conditions remain a mystery, and Siepel hopes technology like ExtRaINSIGHT will help reveal their origins and guide diagnoses and treatments. He also hopes his work will help further illustrate how mutations continue to shape the evolution of the human genome.

Noah Dukler, Mehreen R. Mughal, Ritika Ramani, Yi-Fei Huang & Adam Siepel, Extreme purifying selection against point mutations in the human genome" was published in *Nature Communications* on July 25, 2022.

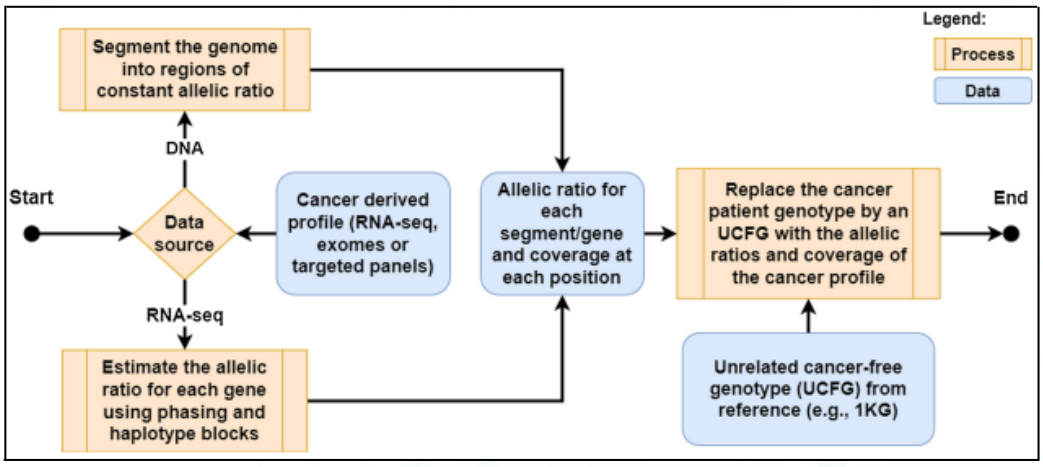
What can race and ethnicity tell us about cancer?

While all populations risk developing cancer in their lifetimes, there are types of the disease whose burden disproportionately falls on certain racial or ethnic groups. Making sense of the molecular makeup of cancer across ancestral groups requires copious amounts of cancer-derived "big data" and accurate knowledge of the patient's genetic ancestry. At present, however, knowing a patient's ancestry often relies on either self-reported data, which often fail to capture complete ancestral information, or on genotyping a patient's DNA from a cancer-free tissue, a practice outside clinical care. **Research Professor Alexander Krasnitz** and his team created a computational framework to determine a patient's ancestry directly from cancer-derived data yielded by various genomic profiling platforms. This method does not require knowledge of the patient's cancer-free genome. The researchers tested their approach in four types of cancer across three different genomic sequencing modalities.

This study demonstrates that the vast amounts of existing cancer data are potentially amenable to studies involving the ancestry of cancer disease without matching cancer-free genomes or patient self-reported ancestry.

They found their method could infer a patient's ancestry with high accuracy compared to the gold standard of determining ancestry by matching a patient's DNA from cancer-free tissue. This study demonstrates that the vast amounts of existing cancer data are potentially amenable to studies involving the ancestry of cancer disease without matching cancer-free genomes or patient self-reported ancestry. By leveraging massive additional data resources, tools developed by the team will make it possible to investigate cancer in all ancestral groups, including those rarely represented in the existing cancer data.

Pascal Belleau, Astrid Deschênes, Nyasha Chambwe, David A. Tuveson & Alexander Krasnitz, "Genetic ancestry inference from cancer-derived molecular data across genomic and transcriptomic platforms" was published in *Cancer Research* on November 9, 2022



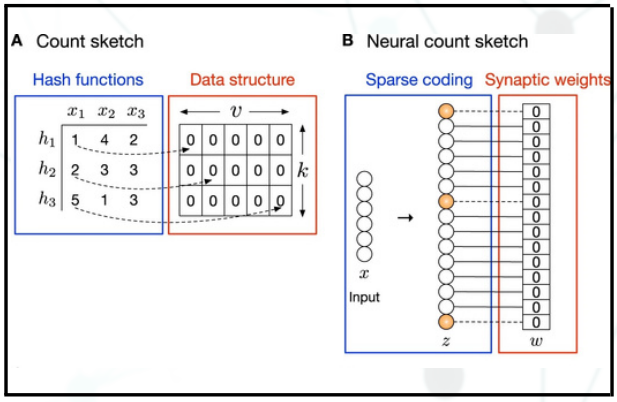
A flow chart of the inference of genetic ancestry.

Even fruit flies count

Imagine sitting in your house when you hear a knock on your front door. You peer through the window and see a stranger outside. Almost immediately, your brain knows you've never encountered this person before. Maybe you feel scared and pretend you're not home. Now, imagine a close friend or family member outside instead of a stranger. Almost as soon as you see them, your brain knows you've encountered this person many times. Our brains keep track of the number of times we experience different stimuli to influence how we respond to a given situation. Even if we don't remember all the details, this count affects whether we avoid or embrace the experience. Evolutionarily, this split-second response could mean the difference between life and death, and scientists are still determining how this process works in our brains. **Associate Professor Saket Navlakha** created a new computational model based on fruit fly brain data to bring us closer to the answer. The fruit fly has about a hundred thousand neurons, and connections at the level of the individual neurons and synapses are known, making it a good model for understanding how the human brain works. In a recently published paper in **Nature Communications**, Navlakha and his team reviewed data from a 2017 study that imaged fruit fly brains exposed to new and familiar odors. Navlakha hypothesized that the insects could distinguish between four categories, something they smelled for the first time, the second time, the third time, and something they had encountered many times before. He created what he calls a "1-2-3-many" count sketch.

Navlakha suspects humans may use a similar, if more complex, counting system and plans to explore this in future research.

Count sketches are used in computer science to provide a quick, approximate tally. For example, YouTube uses a count sketch to track how often a video has been watched, and Navlakha suspects living organisms might do the same thing. By taking inspiration from this classical computer science method, the researchers found that fruit flies use a variant of the count sketch method to track how many times they experience different odors.



The count sketch and corresponding neural circuit implementation.

Navlakha suspects humans may use a similar, if more complex, counting system and plans to explore this in future research. The team would like to see if this model can also explain how this computation may be affected in neurodegenerative disorders like Alzheimer's and Parkinson's disease.

Sanjoy Dasgupta, Daisuke Hattori & Saket Navlakha, " A neural theory for counting memories" was published in *Nature Communications* on October 20, 2022.

Making the most of short-read sequencing data

Advances in DNA sequencing technologies have made it possible to efficiently characterize large segments of individual genomes with the potential to facilitate the next major advancement in medical genetics. In particular, short-read sequencers provide highly accurate reads at a low cost. However, short-read sequencers cannot separate or 'phase' maternally and paternally derived sequence information to understand the relationship between human DNA sequence and phenotype fully. In contrast, long-read sequencers generate data useful for high-quality genome assemblies that can also separate distant heterozygous sites but are error-prone and costly. In a recent study, **Associate Professor Dan Levy** and Research Assistant Professor Siran Li describe novel bench protocols and algorithms to obtain haplotype-phased sequence assemblies with ultra-low errors for regions 10 kb and longer using short reads only. Levy and his team accomplish this by imprinting each template strand from a target region with a dense and unique mutation pattern. The mutation process randomly and independently converts ~50% of cytosines to uracils.

Levy hopes his protocol will enable future genetic studies requiring expensive data from long reads using more cost-effective and accurate short-read sequencing.

Sequencing libraries are made from both mutated and unmutated templates. Using de Bruijn graphs and paired-end read information, the researchers assemble each mutated template and use the unmutated library to correct the mutated bases. Templates are partitioned into two or more haplotypes, and the final haplotypes are assembled and corrected for residual template mutations and PCR errors. With sufficient template coverage, the final assemblies have per-base error rates below 10^{-9}. The group demonstrated this method on a four-member nuclear family, correctly assembling and phasing three genomic intervals, including the highly polymorphic HLA-B gene. Levy hopes his protocol will enable future genetic studies requiring expensive data from long reads using more cost-effective and accurate short-read sequencing.

Siran Li, Sarah Park, Catherine Ye, Cassidy Dankyo, Matthew Wroten, Peter Andrews, Michael Wigler & Dan Levy, "Targeted de novo phasing and long-range assembly by template mutagenesis" was published in *Nucleic Acids Research* on October 14th, 2022.

Complex long assembly of the HLA-B gene.

Ph.D. Training Program

The SCQB is a hub for research, training, and education in the quantitative sciences.

Lead Instructor Justin Kinney.

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 lecture series focuses on various topics in machine learning, algorithms, population genetics, functional genomics, image analysis, and biophysics.

QB students are also encouraged to participate in the **SCQB Mentored Independent Study Program** (p. 12)**.** 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.

COLD SPRING HARBOR SCHOOL OF BIOLOGICAL SCIENCES 2021 DOCTORAL RECIPIENTS

BENJAMIN HARRIS, PHD

Dr. Benjamin Harris conducted his research in Dr. Jesse Gillis' laboratory. His thesis was entitled "Evaluating the functional landscape of Blood and Brains using scRNAseq." Ben is currently a Computational Biologist at Lyell Immunopharma.

SHAINA LU, PHD

Dr. Shaina Lu conducted her research in the laboratories of Drs. Jesse Gillis and Anthony Zador. Her thesis was entitled "The replicability of spatially-resolved transcriptomics for modern neuroscience." She is currently a Bioinformatics Scientist at Eagle Genomics.

KATHRYN O'NEILL, PHD

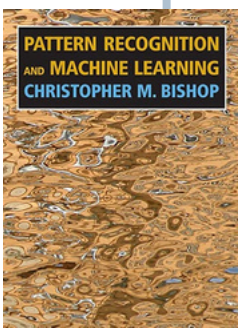
Dr. Kathyrn O'Neill conducted her research in Dr. Molly Gale Hammell's laboratory. Her thesis was entitled "Understanding the contribution of retrotransposon activation to neurodegenerative disease." Kat is currently a CHERPP-T Postdoctoral Fellow at the University of Washington.

Advanced Coursework in Quantitative Biology

Our faculty provide extensive mentoring and support to trainees interested in advanced topics in QB.

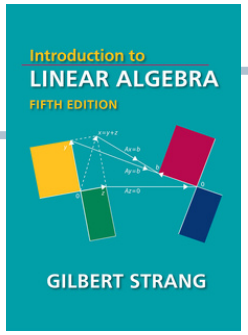
In 2020, faculty members initiated an independent mentored study program to increase QB training in Ph.D. students. The program is done in parallel with their Ph.D. research, where students spend a fraction of their time reading textbooks and then meet with SCQB faculty to discuss their progress.

2020



MACHINE LEARNING

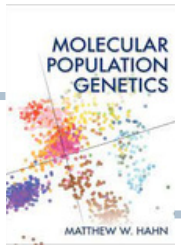
Ziqi (Amber) Tang, Shushan Toneyan, Cole Wunderlich & Xingyu (CiCi) Zhang were jointly mentored by Justin Kinney, Peter Koo, Alexander Krasnitz, David McCandlish & Adam Siepel.



LINEAR ALGEBRA

Ziqi (Amber) Tang & Shushan Toneyan were mentored by David McCandlish.

2021

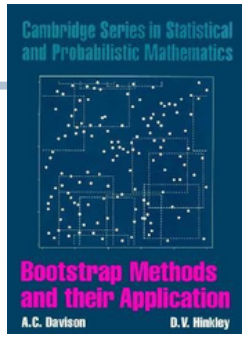


POPULATION GENETICS

Ziyi Mo was mentored by David McCandlish.

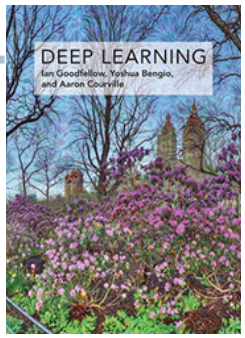
BOOTSTRAP

Michael Passalacqua & Jonathan Werner were mentored by Peter Koo.



DEEP LEARNING

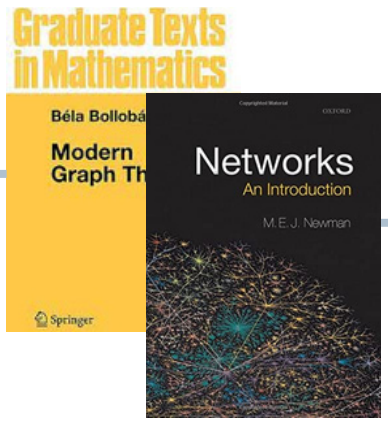
Ziqi (Amber)Tang & Shushan Toneyan were mentored by Justin Kinney.



2022

QUANTITATIVE GENETICS

Shushan Toneyan was mentored by Hannah Meyer.



GRAPH THEORY

Xingyu (CiCi) Zheng was mentored by David McCandlish

CSHL graduate students in SCQB labs are publishing their work in leading scientific journals.

A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph

Hussein A. Hejase,[1,†] Ziyi Mo,[†,1,2] Leonardo Campagna,[3,4] and Adam Siepel[*,1]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[2]School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA
[3]Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, Ithaca, NY, USA
[4]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA
[†]These authors contributed equally to this work.
*Corresponding author: E-mail: asiepel@cshl.edu.
Associate editor: Tal Pupko

Abstract

Detecting signals of selection from genomic data is a central problem in population genetics. Coupling the rich information in the ancestral recombination graph (ARG) with a powerful and scalable deep-learning framework, we developed a novel method to detect and quantify positive selection: Selection Inference using the Ancestral recombination graph (SIA). Built on a Long Short-Term Memory (LSTM) architecture, a particular type of a Recurrent Neural Network (RNN), SIA can be trained to explicitly infer a full range of selection coefficients, as well as the allele frequency trajectory and time of selection onset. We benchmarked SIA extensively on simulations under a European human demographic model, and found that it performs as well or better as some of the best available methods, including state-of-the-art machine-learning and ARG-based methods. In addition, we used SIA to estimate selection coefficients at several loci associated with human phenotypes of interest. SIA detected novel signals of selection particular to the European (CEU) population at the *MC1R* and *ABCC11* loci. In addition, it recapitulated signals of selection at the *LCT* locus and several pigmentation-related genes. Finally, we reanalyzed polymorphism data of a collection of recently radiated southern capuchino seedeater taxa in the genus *Sporophila* to quantify the strength of selection and improved the power of our previous methods to detect partial soft sweeps. Overall, SIA uses deep learning to leverage the ARG and thereby provides new insight into how selective sweeps shape genomic diversity.

Key words: ancestral recombination graph, machine learning, positive selection, selective sweep.

Introduction

Open Access

Mol. Biol. Evol. 39(1):msab332 doi:10.1093/molbev/msab332 Advance Access publication November 22, 2021 1

Machine-learning methods to detect natural selection

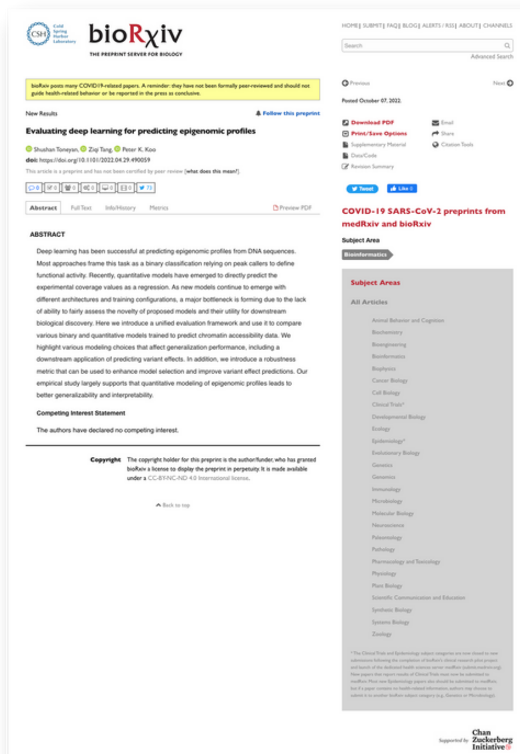
Ziyi Mo is a Ph.D. student in Adam Siepel's lab interested in applying machine-learning methods to detect natural selection in the genome. He has been experimenting with features extracted from ancestral recombination graphs (ARGS) to detect signals of selection. Recently, he co-led a project published in **Molecular Biology and Evolution** to develop an improved machine-learning method called Selection Inference using the ARG (SIA), which refines predictions of selective sweeps and estimation of selection coefficients by taking advantage of a high-dimensional set of features extracted from inferred ARGs.

Hussein A. Hejase, Ziyi Mo, Leonardo Campagna, and Adam Siepel, "A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph" was published in *Molecular Biology and Evolution* on January 7th, 2022.

Deep Learning for Epigenomic Profiling

Shushan Toneyan and **Ziqi (Amber) Tang** are Ph.D. students in Peter Koo's lab. Shushan and Amber use data-driven machine-learning solutions to understand how epigenetics alter gene expression. In a recent study, now accepted in **Nature Machine Intelligence**, they introduced a unified evaluation framework. They used it to compare various binary and quantitative models trained to predict chromatin accessibility data. Their study largely supports the idea that quantitative modeling of epigenomic profiles leads to better generalizability and interpretability.

Shushan Toneyan, Ziqi Tang & Peter Koo, "Evaluating deep learning for predicting epigenomic profiles" was posted to *bioRxiv* on October 7th, 2022, and was accepted for publication in *Nature Machine Intelligence*.



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Evaluating deep learning for predicting epigenomic profiles

Shushan Toneyan, Ziqi Tang, Peter K. Koo
doi: https://doi.org/10.1101/2022.04.29.490059

Posted October 07, 2022.

ABSTRACT

Deep learning has been successful at predicting epigenomic profiles from DNA sequences. Most approaches frame this task as a binary classification relying on peak callers to define functional activity. Recently, quantitative models have emerged to directly predict the experimental coverage values as a regression. As new models continue to emerge with different architectures and training configurations, a major bottleneck is forming due to the lack of ability to fairly assess the novelty of proposed models and their utility for downstream biological discovery. Here we introduce a unified evaluation framework and use it to compare various binary and quantitative models trained to predict chromatin accessibility data. We highlight various modeling choices that affect generalization performance, including a downstream application of predicting variant effects. In addition, we introduce a robustness metric that can be used to enhance model selection and improve variant effect predictions. Our empirical study largely supports that quantitative modeling of epigenomic profiles leads to better generalizability and interpretability.

Competing Interest Statement

The authors have declared no competing interest.

Copyright The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

EDUCATION & TRAINING

Postdoctoral Training in Quantitative Biology

The SCQB supports postdoctoral fellows' career development and research training in a collaborative, interdisciplinary environment.

SCQB postdoctoral fellows are part of an interdisciplinary environment where they receive training in innovative modeling, algorithmic, and machine-learning methods, drawing broadly from mathematics, computer science, and physics techniques to address open problems in biology. Many of our postdocs finish their appointments within two years and go on to exciting positions in academia and industry.

We offer an innovative funding opportunity for postdoctoral research and training, known as the **Interdisciplinary Scholars in Experimental and Quantitative Biology (ISEQB)**. Launched in 2018, this program is designed to help recruit new postdocs or fund existing CSHL postdocs interested in cross-training in experimental and computational research laboratories (p. 25).

2022 SCQB ALUMNI

MAHDI KOOSHKBAGHI, PHD

Dr. Mahdi Kooshkbaghi, formerly a postdoc in Justin Kinney's laboratory, is now a Machine Learning Engineer at Estee Lauder.

ANTONIO MAJDANDZI, PHD

Dr. Antonio Majdandzi, formerly a postdoc in Peter Koo's laboratory, is now a Machine Learning Researcher at Bloomberg.

MEHREEN MUGHAL, PHD

Dr. Mehreen Mughal, formerly a postdoc in Adam Siepel's laboratory, is now a Scientist in Population Genetics at Variant Bio in Seattle, Washington.

AMMAR TEHREEN, PHD

Dr. Ammar Tehreen, formerly a postdoc in Justin Kinney's laboratory, is now a Scientist at Regeneron Pharmaceuticals in Tarrytown, New York.

ROHIT TRIPATHY, PHD

Dr. Rohit Tripathy, formerly a postdoc in Peter Koo's laboratory, is now an Associate Computational Scientist at The Jackson Laboratory.

Meetings and Affiliations

Our faculty maintain close collaborative ties across CSHL and other NY area groups and organize relevant QB meetings and conferences.

CSHL Cancer Center

The **CSHL Cancer Center** now features QB as a major theme and includes 80% of our core faculty as members. Research Professor **Alexander Kraznitz** (p. 8) and Fellow **Hannah Meyer** partnered with traditional cancer laboratories, producing several publications and preprints this year. In particular, Hannah Meyer, in collaboration with Assistant Professor Tobias Janowitz, used genome-wide association and structural equation modeling to predict the failure of cancer immunotherapy. In a separate preprint, Hannah and Assistant Professor Semir Beyaz studied how CRISPR/Cas9 modeling influences gain-of-function mutations of oncogenes to drive distinct subtypes of liver cancer.

Flatiron Institute

Associate Professor **Peter Koo** and Assistant Professor **Justin Kinney** co-organized a one-day meeting with Michael Shelley, the Center for Computational (CCB) Director at the Flatiron Institute.

The **SCQB CCB Show & Tell** took place on October 12, 2022, at the Flatiron Institute in Manhattan and comprised ten faculty talks and a trainee poster session. This meeting provided a unique opportunity for trainees and faculty alike to meet and exchange ideas in a friendly, intimate environment. Speakers included SCQB faculty **Justin Kinney**, **Peter Koo**, **David McCandlish**, **Saket Navlakha**, **Adam Siepel**, and CCB group leaders Sonya Hanson, Pilar Cossio, Julia Koehler Leman, Olga Troyankaya, Michael Shelley, and Stas Shvartsman.

New York Genome Center (NYGC)

Professor **Adam Siepel** continues to co-lead the **Population Genomics Working Group** with David Knowles, Ph.D. of the NYGC. At the same time, Associate Professor **Ivan Iossifov** maintains his joint appointment as a core faculty member at the NYGC, where his research focuses on large autism sequencing projects. He presented an update on his autism work to the Simons Foundation Autism Research Initiative (SFARI) team at their New York office in March of this year.

Meetings Organized by SCQB Faculty

- **Distributed Computing Perspectives on Theoretical Immunology Working Group**, Santa Fe Institute, Santa Fe, NM, June 9, 2022, Co-organized by Saket Navlakha
- **FASEB Mobile DNA**, Dublin, Ireland, June 4-9, 2022, Co-organized by Molly Gale Hammell
- **GS3: Mutational Biases and Adaptation**, Virtual Conference, August 2, 2022, Co-organized by David McCandlish
- **SCQB CCB Show & Tell**, Flatiron Institute, New York, NY, October 12, 2022, Co-organized by Justin Kinney and Peter Koo

Fellow Hannah Meyer's research featured on WNYC's Radiolab

In August of this year, Fellow **Hannah Meyer** was featured on **Radiolab**, a Peabody award-winning radio show focusing on scientific, philosophical, and political topics, for her work on the thymus. The episode "Inner Workings" explores one of the most philosophical places in the universe: the thymus, an organ that knows what is you and what is not. Naturally, the interviewers were interested in speaking with Hannah Meyer, whose research focuses on how the thymus generates and selects a highly variable yet specific T cell repertoire that discriminates between healthy and non-healthy self and dangerous non-self antigens. Hannah discusses how the thymus acts as the biological training ground where the body learns to protect itself from outside invaders like bacteria and coronaviruses.

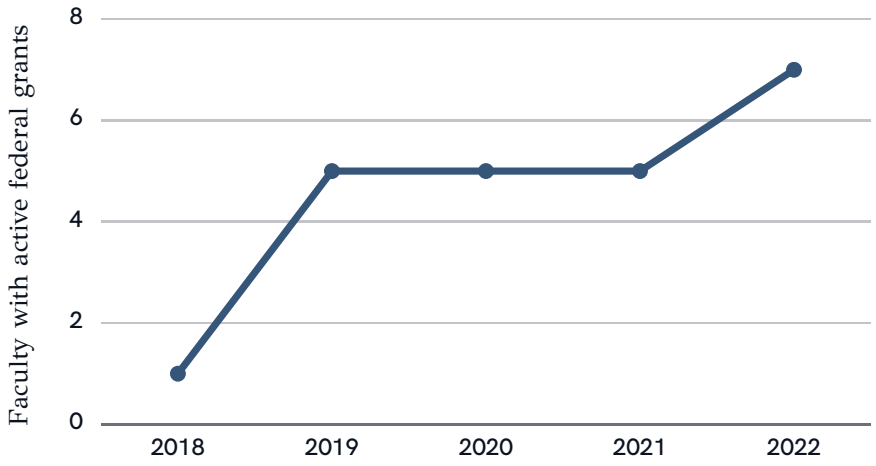
Associate Professor Molly Gale Hammell commits to open science through work with the CZI NDCN

In 2018, Associate Professor **Molly Gale Hammell** was awarded the Chan Zuckerberg Initiative (CZI) Ben Barres Early Career Acceleration Award for her work using machine learning software to study amyotrophic lateral sclerosis. Molly is part of the **CZI Neurodegeneration Challenge Network (NDCN)**, which brings together experimental scientists from diverse research fields, along with computational biologists and physicians, to understand the fundamental biology of neurodegenerative disorders. In 2021, Molly was awarded the CZI NDCN Network Annual Meeting Open Science Award, and she currently leads the **CZI Single-Cell Analysis Working Group** as part of the CZI NDCN Network.

Graduate Student Shushan Toneyan awarded NVIDIA GPU Grant

NVIDIA, a technology company known for designing and manufacturing graphics processing units (GPUs), believes academic institutions are at the forefront of nurturing the next generation in the emerging technologies of accelerated computing, data science, and AI. The **NVIDIA Academic Hardware Grant Program** was developed to advance research with access to world-class computing resources. This year, Ph.D. student **Shushan Toneyan** was awarded a NVIDIA GPU grant to advance her research.

Faculty members of the SCQB continue to secure federal funding for their state-of-the-art research



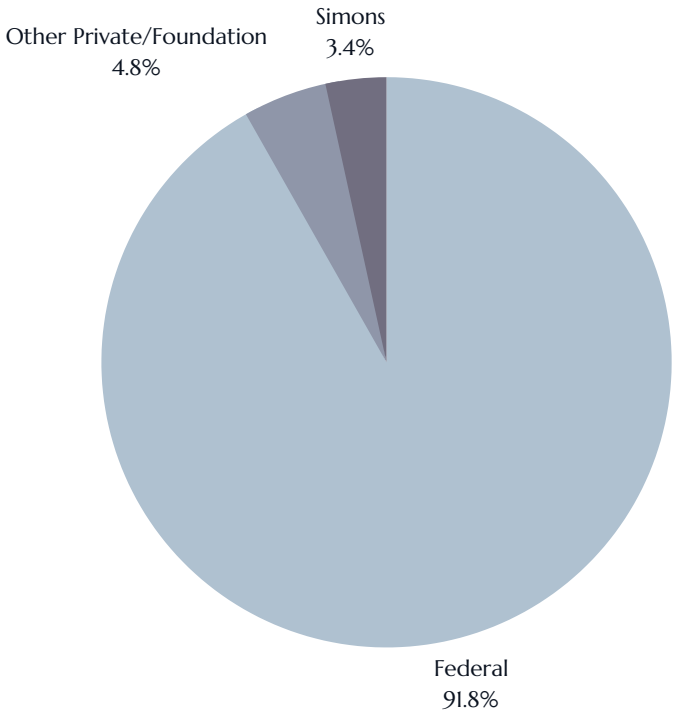
70%

Core faculty hold R01 funding in 2022

The number of faculty holding active federal grants has steadily increased since 2018, where seven of ten core faculty members now hold at least one NIH R01 or equivalent grant.

2022 FUNDING BREAKDOWN

The federal government now funds over 90% of the research at the SCQB.



$9.2M* Total funding awarded from the National Institutes of Health in 2022.

*$4,782,921 Direct/$4,370,144 Indirect

Grant funded projects and their rationale.

1R01HG011787-01A1
"A unified quantitative modeling strategy for multiplex assays of variant effect"
National Human Genome Research Institute
Justin Kinney (Contact PI)

A diverse class of experimental methods known as multiplex assays of variant effect (MAVEs) is revolutionizing the scientific understanding of how genomes encode information. **Associate Professor Justin Kinney** will develop a unified computational framework for analyzing the large datasets that MAVEs produce. These methods will provide unprecedented quantitative insight into how genomic DNA encodes biologically critical information, thus advancing the clinical understanding of human genetic variation and accelerating the development of treatments for various diseases.

1R01HG012131-01A1
"Interpretable computational models of functional genomics data"
National Human Genome Research Institute
Peter Koo (Contact PI)

The complex coordination of cis-regulatory elements (CREs), the so-called cis-regulatory code, regulates many biological processes, including transcriptional regulation and alternative splicing. **Assistant Professor Peter Koo** will develop machine learning-based methods to decipher this code by quantitatively characterizing higher-order interactions within CREs and between multiple CREs using functional genomics data. His work will directly apply to fundamental problems in basic biological science and the understanding and treatment of human diseases.

1R01AI167862-01
"Dissecting dynamic genetic effects from thymus development to immune-mediated disease"
National Institute of Allergy and Infectious Diseases
Hannah Meyer (Contact PI)

Flaws in the immune system lead to immunodeficiency and autoimmune diseases. **CSHL Fellow Hannah Meyer** proposes combining quantitative genetics, experimental methods, and clinical samples to study the mechanisms of these flaws and gain insights into how they impact immune-mediated diseases.

APPENDIX

PUBLICATIONS

1 Abuhashem A, Chivu AG, Zhao Y, Rice EJ, **Siepel A**, Danko CG, Hadjantonakis AK. RNA Pol II pausing facilitates phased pluripotency transitions by buffering transcription. *Genes Dev*., 36(13-14):770–89, 2022.

2 Baslan T, Morris JP 4th, Zhao Z, Reyes J, Ho YJ, Tsanov KM, Bermeo J, Tian S, Zhang S, Askan G, Yavas A, Lecomte N, Erakky A, Varghese AM, Zhang A, Kendall J, Ghiban E, Chorbadjiev L, Wu J, Dimitrova N, Chadalavada K, Nanjangud GJ, Bandlamudi C, Gong Y, Donoghue MTA, Socci ND, **Krasnitz A**, Notta F, Leach SD, Iacobuzio-Donahue CA, Lowe SW. Ordered and deterministic cancer genome evolution after p53 loss. *Nature*. 608(7924):795-802, 2022.

3 Belleau P, Deschênes A, Chambwe N, Tuveson D, **Krasnitz A**. Genetic ancestry inference from cancer-derived molecular data across genomic and transcriptomic platforms. *Cancer Res*. [Internet], 2022.

4 Bhattacharya N, Thomas N, Rao R, Dauparas J, **Koo PK**, Baker D, Song YS, Ovchinnikov S. Interpreting Potts and Transformer Protein Models Through the Lens of Simplified Attention. *Pac Symp Biocomput*., 27:34-45, 2022.

5 Brown AL, Wilkins OG, Keuss MJ, Hill SE, Zanovello M, Lee WC, Bampton A, Lee FCY, Masino L, Qi YA, Bryce-Smith S, Gatt A, Hallegger M, Fagegaltier D, Phatnani H; NYGC ALS Consortium, **Hammell MG**, Newcombe J, Gustavsson EK, Seddighi S, Reyes JF, Coon SL, Ramos D, Schiavo G, Fisher EMC, Raj T, Secrier M, Lashley T, Ule J, Buratti E, Humphrey J, Ward ME, Fratta P. TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature*, 131-137, 2022

6 Campagna L, Mo Z, **Siepel A**, Uy JAC. Selective sweeps on different pigmentation genes mediate convergent evolution of island melanism in two incipient bird species. *PLoS Genet*., 18(11):e1010474, 2022.

7 Cano AV, Rozhoňová H, Stoltzfus A, **McCandlish DM**, Payne JL. Mutation bias shapes the spectrum of adaptive substitutions. *Proc Natl Acad Sci U S A*. 2022 Feb 15;119(7):e2119720119, 2022.

8 Carter JA, Strömich L, Peacey M, Chapin SR, Velten L, Steinmetz LM, Brors B, Pinto S, **Meyer HV**. Transcriptomic diversity in human medullary thymic epithelial cells. *Nat Commun*. 13(1):4296, 2022.

9 Dasgupta S, Hattori D, **Navlakha S**. A neural theory for counting memories. *Nat Commun*.13(1):5961, 2022.

10 Dukler N, Mughal MR, Ramani R, Huang YF, **Siepel A**. Extreme purifying selection against point mutations in the human genome. *Nat Commun*. 13(1):4312, 2022.

11 Gao Y, He XY,…,Toneyan S,…**Koo PK**,…Vakoc C. ETV6 dependency in Ewing sarcaoma by antagonism of EWS-FLI1-mediated enhancer activation. *Nat. Cell Biol*. [Accepted], 2022.

12 Garza R, Sharma Y, Atacho D, Hamdeh SA, Jönsson M, Ingelsson M, Jern P, **Hammell MG**, Englund E, Jakobsson J, Marklund N. Single-cell transcriptomics of resected human traumatic brain injury tissues reveals acute activation of endogenous retroviruses in oligodendrocytes. *bioRxiv*, 506982 [Preprint], 2022.

13 Hejase HA, Mo Z, Campagna L, **Siepel A**. A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. *Mol Biol Evol*. 39(1):msab332, 2022.

14 Kaczmarzyk, JR, Kurc, TM, Abousamra S, Gupta R, Saltz JH, **Koo PK**. Evaluating histopathology transfer learning with ChampKit. *arXiv*, 2206.06862 [Preprint], 2022.

15 Kawaguchi RK, Tang Z, Fischer S, Rajesh C, Tripathy R, **Koo PK**, Gillis J. Learning single-cell chromatin accessibility profiles using meta-analytic marker genes. *bioRxiv*, 438068 [Preprint], 2022.

16 Kleeman SO, Demestichas B, Mourikis N, Loiero D, Ferrer M, Bankier S, Riazat-Kesh YJRA , Lee H, Chantzichristos D, Regan C, Preall J, Sinha S, Rosin N, Yipp B, de Almeida, LGN, Biernaskie J, Dufour A, Tober-Lau P, Ruusalepp A, Bjorkegren JLM, Ralser M, Kurth F, Demechev V, Heywood T, Gao Q, Johannsson G, Koelzer VH, Walker BR, **Meyer HV**, Janowitz T. Cystatin C is glucocorticoid-responsive, directs recruitment of Trem2+ macrophages and predicts failure of cancer immunotherapy. *medRxiv*, 21261668 [Preprint], 2022.

17 Kleeman SO, Cordioli M, Timmers PRHJ, Khan A, Tober-Lau P, Kurth F, Demichev V, **Meyer HV**, Wilson JF, Ralser M, Kiryluk K, Ganna A, Baillie K, Janowitz T. Cystatin C is associated with adverse COVID-19 outcomes in diverse populations. *iScience*. 25(10):105040 [Epub], 2022.

18 Mou H, Yue J, Jin Y, Wang Z, Gao Y. Janowitz T, **Meyer HV**, Kucukural A, Wilkinson JE, Ozata DM, Beyaz S. CRISPR-induced exon skipping of b-catenin reveals tumorigenic mutants driving distinct subtypes of liver cancer. *bioRxiv*, 485965 [Preprint], 2022.

19 Lee N, Tang Z, Toneyan S, **Koo PK**. Evolution-inspired augmentations improve deep learning for regulatory genomics *bioRxiv*, 515117 [Preprint], 2022.

20 Li S, Park S, Ye C, Danyko C, Wroten M, Andrews P, Wigler M, **Levy D**. Targeted de novo phasing and long-range assembly by template mutagenesis. *Nucleic Acids Res*., 50(18):e103, 2022.

21 Majdandzic A, Rajesh C, **Koo PK**.Correcting gradient-based interpretations of deep neural networks for genomics. *bioRxiv*, 490102 [Preprint], 2022.

22 Pukhrambam C, Molodtsov V, Kooshkbaghi M, Tareen A, Vu H, Skalenko KS, Su M, Yin Z, Winkelman JT, **Kinney JB**, Ebright RH, Nickels BE. Structural and mechanistic basis of σ-dependent transcriptional pausing. *Proc Natl Acad Sci U S A*., 119(23):e2201301119, 2022.

23 Tareen A, Kooshkbaghi M, Posfai A, Ireland WT, McCandlish DM, **Kinney JB**. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biol*., 23(1):98, 2022.

24 Toneyan S, Tang Z, Koo PK.Evaluating deep learning for predicting epigenomic profiles. Nat. Mach. Intell. bioRxiv, 490059 [Preprint, Accepted in *Nat. Mach. Intell.*], 2022.

25 Thanaj M, Mielke J, McGurk KA, Bai W, Savioli N, de Marvao A, **Meyer HV**, Zeng L, Sohler F, Lumbers RT, Wilkins MR, Ware JS, Bender C, Rueckert D, MacNamara A, Freitag DF, O'Regan DP. Genetic and environmental determinants of diastolic heart function. *Nat Cardiovasc Res*. (4):361-371 [Epub] 2022.

26 Suen JY, **Navlakha S**. A feedback control principle common to several biological and engineered systems. *J R Soc Interface*. 19(188):20210711 [Epub], 2022.

27 Šušnjar U, Škrabar N, Brown AL, Abbassi Y, Phatnani H; NYGC ALS Consortium, **Hammell MG**, Cortese A, Cereda C, Bugiardini E, Cardani R, Meola G, Ripolone M, Moggio M, Romano M, Secrier M, Fratta P, Buratti E. Cell environment shapes TDP-43 function with implications in neuronal and muscle disease. *Commun Biol*. 5(1):314, 2022.

28 Wang Z, Moffitt AB, Andrews P, Wigler M, **Levy D**. Accurate measurement of microsatellite length by disrupting its tandem repeat structure. *Nucleic Acids Res*. gkac723 [Epub ahead of print], 2022.

29 Weinstein JY, Aldaravi CMG., Lipsh-Sokolik R, Hoch SY, Liebermann D, Nevo R, Weissman H, Petrovich-Kopitman E, Margulies D, Ivankov D, **McCandlish D**. Designed active-site library reveals thousands of functional GFP variants. *bioRxiv* 511732 [Preprint], 2022

30 Wu Y, Johnson L, Song B, Romay C, Stitzer M, **Siepel A**, Buckler E, Scheben A. A multiple alignment workflow shows the effect of repeat masking and parameter tuning on alignment in plants. *Plant Genome*. (2):e20204 [Epub], 2022.

31 Zhao Y, Liu L, **Siepel A**. Model-based characterization of the equilibrium dynamics of transcription initiation and promoter-proximal pausing in human cells. *bioRxiv*, 512929 [Preprint], 2022.

32 Zhou J, Wong MS, Chen WC, Krainer AR, **Kinney JB**, **McCandlish DM**. Higher-order epistasis and phenotypic prediction. *Proc Natl Acad Sci U S A*., 119(39):e2204233119 [Epub], 2022.

New External Funding in 2022

PI	TOTAL AWARD[1]	YEARS	FUNDING SOURCE	PROJECT TITLE
Iossifov, I.	$343,325	2022-2023	The Simons Foundation	Genotype and phenotype in families (GPF) systems
Kinney J. McCandlish. D.[2]	$3,818,348 $301,279	2022-2027	National Institutes of Health	A unified quantitative modeling strategy for multiplex assays of variant effect
Koo, P. Kinney, J.[2] McCandlish, D.[2]	$2,063,031 $43,369 $38,899	2022-2027	National Institutes of Health	Interpretable computational models of functional genomic data
McCombie, WR. Krasnitz, A.[2]	$315, 103 $131, 918	2021-2023	CSHL-Northwell Health	Biological basis of differential colon cancer mortality in African Americans (AAs) vs. European Americans (EAs): A pilot study
Meyer, H.	$2,888,139	2022-2027	National Institutes of Health	Dissecting dynamic genetic effects from thymus development to immune mediated disease

[1]Total award, including indirect costs

[2]PI designated this amount to the individual QB investigator

Laboratory Membership

FACULTY	LAB MEMBER	POSITION	YEAR STARTED
Molly Gale Hammell	Talitha Forcier	Postdoctoral Fellow	2018
	Oliver Tam	Computational Science Analyst	2017
	Karthick Natarjan	Post Doc Computational	2021
	Regina Shaw	Research Technician	2016
	Cole Wunderlich	Graduate Student, CSHL	2019
	Craig Marshall	Graduate Student, SBU	2019
	Isolbel Bolger	Graduate Student, SBU	2020
Ivan Iossifov	Yoon-Ha Lee	Research Investigator	2015
	Steven Marks	Computational Science Developer	2017
Justin Kinney	Taehoon Ha	Biostatistician	2020
	Evam Seitz	Post Doc Computational	2022
	Deborah Tenenbaum	Post Doc Computational	2022
	Andalus Ayaz	Research Technician	2016
Peter Koo	Antonio Majdandzic	Post Doc Computational	2020
	Jakub Kaczmarzyk	Graduate Student, Visiting	2021
	Chandana Rajesh	Graduate Student, SBU	2022
	Ziqi (Amber) Tang	Graduate Student, CSHL	2019
	Shushan Toneyan	Graduate Student, CSHL	2019
Alexander Krasnitz	Nissim Ranade	Postdoc Computational	2018
	Pascal Belleau	Post Doc Computational	2017
	Taimour Baslan	Collaborative Scientist	2016
Dan Levy	Zhezhen Yu	Graduate Student, SBU	2017
David McCandlish	Carlos Martí Gomez	Post Doc Computational	2021
	Michele Avella	Graduate Student Visiting	2022
	Julian Klug	Visiting Scientist	2022
	Bryan Gitschlag	Post Doc Computational	2021
Hannah Meyer, CSHL Fellow	Salomé Carcy	Graduate Student, CSHL	2020
	Joshua Torres	Research Technician	2022
	Sarah Chapin	Computational Science Developer	2020
	Kadir Ozler	Research Technician	2021
	Yong Lin	Post Doc Computational	2022

Core Faculty

FACULTY	LAB MEMBER	POSITION	YEAR STARTED
Saket Navlakha	Jalaj Bhandari	Post Doc Computational	2021
	Xingyu (CiCi) Zheng	Graduate Student, CSHL	2020
	Yang Shen	Post Doc Computational	2019
	Matthew Venezia	Undergraduate, Intern	2022
Adam Siepel	Luiz Machado de Oliveira	Graduate Student, SBU	2022
	Armin Scheben	Post Doc Computational	2019
	Alexander (Xander) Xue	Post Doc Computational	2018
	Yixin Zhao	Post Doc Computational	2018
	Lingjie Liu	Graduate Student, SBU	2020
	Ziyi Mo	Graduate Student, CSHL	2018
	Rebecca Hassett	Computational Science Developer	2022

Associated Members

NAME	POSITION
Alexei Koulakov	Professor
W. Richard McCombie	Professor
Partha Mitra	Professor
Jesse Gillis	Associate Professor
Alexander Dobin	Assistant Professor
Tatiana Engel	Assistant Professor
Doreen Ware	Assistant

Administrative Support

NAME	POSITION	YEAR STARTED
Katherine Brenner	QB Science Manager	2018
Susan Fredericks	Assistant to the Chair and Sr. Scientific Administrator	2022
Idee Mallardi	Sr. Scientific Adminisrtator	2016

Interdisciplinary Scholars in Quantitative and Experimental Biology (ISEQB)Program

AWARD YEAR	SCHOLAR	PROJECT TITLE	QB MENTOR	EXPERIMENTAL MENTOR
2017-2021	**Sze 'Mandy' Wong**	Massively parallel reporter assays to decode RNA splicing	Justin Kinney	Adrian Krainer
2018-2022	**Wei-Chia Chen**	New statistical methods for survival analysis on small clinical data	Justin Kinney	Robert Maki
2018-2022	**James Roach**	Contributions of excitatory and inhibitory neurons to decision computations	Tatiana Engel	Anne Churchland
2019	**Walter Bast**	Towards understanding the functional bases of odor perception: quantifying the input-output transfer function of the olfactory bulb	Alexei Koulakov	Florin Albeanu
2022	**Amitava Banerjee**	Compressed sensing in the immune system	Saket Navlakha	Hannah Meyer

Cold Spring Harbor Laboratory
Simons Center for Quantitative Biology

One Bungtown Road
Cold Spring Harbor, NY 11724

cshl.edu/research/quantitative -biology/

Cold Spring Harbor Laboratory
Simons Center for Quantitative Biology

SCQB@cshl.edu

facebook.com/cshlnews/

linkedin..com/school/cold-spring-harbor-laboratory

@CSHL