



Cold Spring Harbor Laboratory

SIMONS CENTER FOR QUANTITATIVE BIOLOGY

2018 ANNUAL REPORT



LETTER FROM THE CHAIR

EVERY YEAR DURING THE OVERCAST and chilly days of late December, I find a measure of brightness and warmth in reflecting on the past year's progress at Cold Spring Harbor Laboratory's Simons Center for Quantitative Biology. 2018 was a particularly eventful year at the SCQB. The Center has now grown to eight principal investigators and nearly 50 people altogether, not including associated investigators and their research groups (see p. 3). The faculty now includes Associate Professor Molly Hammell, who was formerly assigned to the Genomics program at CSHL but, to our delight, is now officially recognized as a QB faculty member. A



Adam Siepel

search is underway for up to two additional faculty members and we are optimistic about several outstanding candidates scheduled to interview in early 2019. We are also pleased to welcome a new SCQB Fellow, Dr. Hannah Meyer, who will be arriving from the European Bioinformatics Institute in March, 2019. Another key addition this year was Katie Brenner, who has taken the jack-of-all-trades position of "QB Science Manager". Since her arrival in March, 2018, Katie has already made a number of important contributions to the Center, including overhauling our website (www.cshl.edu/scqb), helping to craft several grant proposals and journal papers, and taking responsibility for the improved content and appearance of this report. Together with QB postdocs and their faculty advisors, Katie also helped launch a new graduate course in machine learning, which will be organized around an online course from Coursera but will be tutored locally by postdocs Ammar Tareen and Noah Dukler. Remarkably, more than 40 students, postdocs, and staff members from across CSHL have already registered for this course. Another new Laboratory-wide initiative organized by the SCQB is the Interdisciplinary Scholars in Experimental and Quantitative Biology program, which enables postdoctoral scholars to train jointly with members of the SCQB and experimental biologists at CSHL. We have already supported three interdisciplinary postdocs through this program, one of whom, Mandy Wong, is highlighted in this year's report (p. 6). Research activities at the SCQB continue to move forward at full

throttle, with 28 new publications and critical advances in all of our major thematic areas: gene regulation, evolutionary genomics, genomic disease research, and genomic technology development (p. 3; see highlighted projects on pp. 4–5). In 2018, these research advances helped to attract a total of nearly 7 million dollars in new grants, including a prestigious early-career award to Molly Hammell from the Chan Zuckerberg Initiative (p. 8). In addition to these awards, associated faculty member Tatiana Engel of neuroscience, who has been heavily involved with the SCQB, received a major grant from the NIH's BRAIN initiative to fund her research in computational neuroscience (p. 8). We did not have a meeting of our External Advisory Committee this year, but we were pleased to welcome Prof. Molly Przeworski of Columbia University to the EAC. Dr. Przeworski replaces Prof. Dame Janet Thornton of the European Bioinformatics Institute, whom we thank for her generous service over the past three years. The next meeting of the EAC will be in mid-2019. Finally, perhaps one of the best measures of success of a research institute is the quality of its PhD graduates, and we were proud to have three excellent, newly minted PhDs emerge from the SCQB this year: Drs. Noah Dukler, Talitha Forcier, and Brad Gulko (p. 7). In addition, postdoctoral associate Yifei Huang will finish his highly productive stint at the SCQB in December and begin soon afterward as an Assistant Professor in Biology at Penn State University (p. 8). Altogether, 2018 has been an exceptionally active and productive year at the Center, and we anticipate many more exciting developments in 2019.

Best Wishes for the New Year,

Adam Siepel, PhD, Chair
December 30, 2018

<i>Overview.....</i>	3	<i>Education and Outreach.....</i>	7
<i>Research Highlights.....</i>	4	<i>Awards and Recognition.....</i>	8
<i>Collaborations.....</i>	6		

OVERVIEW

THE SIMONS CENTER FOR QUANTITATIVE BIOLOGY IS Cold Spring Harbor Laboratory's home for mathematical, computational, and theoretical research in biology. Over the last 10 years, we have grown from the idea of a program to a group of nearly 50 scientists and staff 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 is permeated by four major themes: **Gene Regulation, Evolutionary Genomics, Genomic Disease Research, and Genomic Technology Development.**

GENE REGULATION

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

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, often in collaboration with clinical oncologists. 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.

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.

MEMBERS OF THE SCQB published research articles in a diverse collection of leading scientific journals in 2018. While these publications include many important findings, we've chosen to highlight four studies that are representative of our 2018 research efforts.

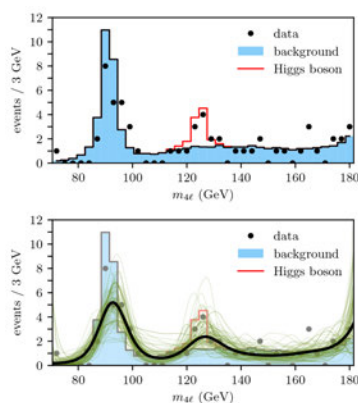
The big problem of small data: A new approach



Justin Kinney

Dealing with small datasets is a fundamental part of doing science, yet quantifying certainty within these datasets has been difficult due to the assumptions that common statistical methods make. Before the computer age when these standard methods were developed, assumptions like these were necessary, but can lead to imprecise approximations when it comes to small data sets. Justin Kinney and his team have crafted a modern computational approach called Density Estimation using Field Theory, or DEFT, that fixes these shortcomings. DEFT is freely available via an open source package called SUFTware. In their recent paper, published in *Physical Review Letters*, Kinney's lab demonstrates DEFT on two datasets: national health statistics compiled by the World Health Organization, and traces of subatomic particles used by physicists at the Large Hadron Collider to reveal the existences of the Higgs boson particle. Kinney plans to adapt DEFT to problems in survival analysis which are widely used in clinical trials.

Wei-Chia Chen, Ammar Tareen, and Justin B. Kinney, "Density estimation on small datasets" was published in *Physical Review Letters* October 18, 2018



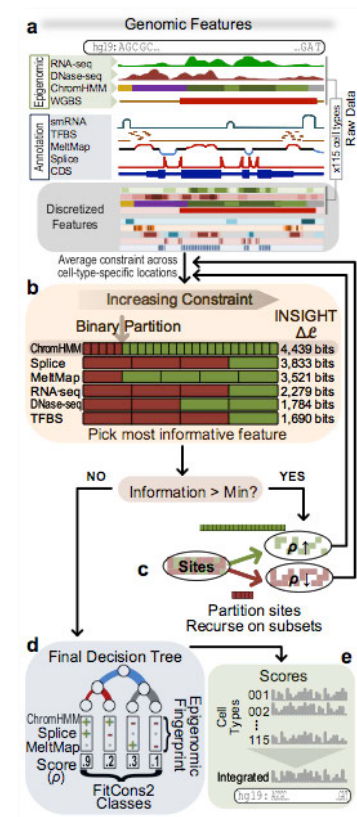
Top: Number of Higgs Boson particle events expected based on Standard Model simulations. Bottom: DEFT was used to smoothly predict (black) how many 4-lepton decay events were indicators of a true Higgs Boson event within a margin of uncertainty (green).

How much are we learning? Natural selection is science's best critic



Adam Siepel

Only about one percent of the human genome accounts for the genes that make the proteins our bodies need to grow and function, but roughly five percent has remained conserved over millions of years of evolution. Therefore, at least about four percent of the genome must be critically important for other, largely unknown reasons. To solve the mystery of the four percent, scientists have spent years developing new technologies to measure various properties of the genome and the proteins that interact with it, collectively known as the epigenome. Current epigenomic measurements, however, provide only indirect indications of biological function, and researchers disagree on how informative they are. Professor Adam Siepel and recent PhD graduate Brad Gulko wanted to let evolution do the work of revealing which parts of the genome are important and how much can be learned about genomic function from each epigenomic data set. They devised a mathematical model to measure the strength of natural selection separately for many different combinations of epigenomic marks, using genome sequences from both modern humans and nonhuman primates. Their model provided a score for every position in the genome indicating how important it is for survival, or more precisely, how likely a mutation at that position would be to reduce a person's "fitness". They



(a) Epigenomic features are arranged along the human genome sequence. (b,c) The FitCons2 algorithm repeatedly partitions the genome by choosing the features that most decreases the "entropy" under a probabilistic evolutionary model. This is the feature that offers the most "information". (d) The end results is a decision tree for clustering genomic sites by their epigenomic "fingerprints" and an estimate of the fitness consequences of a mutation at every genomic position. (e) These scores are mapped back to the genome for each cell type.

released these fitness consequence, or “FitCons”, maps for 15 different human cell types for use by other researchers. Importantly, their model also provided an overall measure of how much “information” each epigenomic data type provides about genomic function, as measured using evolution. Siepel and Gulko argued that most of the apparent information in the genome is actually just “noise” due to mutations at unimportant positions, and that the amount of evolutionary relevant “signal” in the genome is quite small, measuring only a few megabytes – similar in size to a typical email attachment or smartphone photo.

Brad Gulko and Adam Siepel, “An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness” was published in *Nature Genetics* December 17, 2018

De novo mutations diminish motor skills in children with autism



Ivan Iossifov

In individuals with Autism Spectrum Disorder (ASD), de novo mutations have previously been shown to be significantly correlated with lower IQ, but not with core characteristics of ASD such as deficits in social communication and interaction, and restricted interests and repetitive patterns of behavior. Ivan Iossifov and colleagues have extended these findings by demonstrating in the Simons Complex Collection that damaging de novo mutations in ASD individuals are also significantly correlated with measures of impaired motor skills. Iossifov and colleagues found that IQ and motor skills are distinctly associated with damaging mutations and, in particular, that motor skills are a more sensitive indicator of mutational severity than is IQ, as judged by mutational type and target gene. They have used this finding to propose a combined classification of phenotypic severity: mild (little impairment of either), moderate (impairment mainly to motor skills) and severe (impairment of both IQ and motor skills).

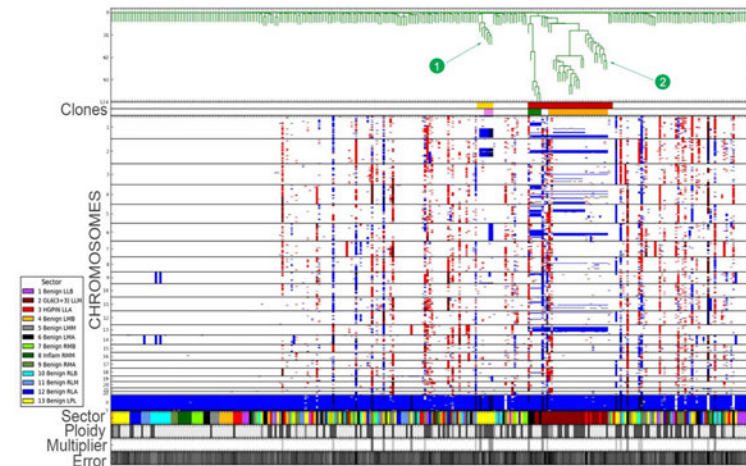
Buja A et al, “Damaging de novo mutations diminish motor skills in children on the autism spectrum” was published in *PNAS* February 20, 2018

Single cell genomics as a diagnostic tool for cancer



Alexander Krasnitz

Prostate cancer is the third most common cancer type among Americans. While most cases are not lethal, 21,000 men will die from prostate cancer this year because it is so common. Alexander Krasnitz and colleagues recently published encouraging results from a pilot study testing a new way



Screen shot from the single cell genomic viewer (SCGV) displaying results from diagnostic prostate biopsy analysis. The genomic profiles of several hundred prostate cells samples from an individual's 13 biopsy cores are arrayed in columns. The upper most section shows the phylogenetic trees (green) that reflect each cell's copy-number profile. The arrows indicate the location of two clones and subclones in the biology that have a strong signal indicating the presence of cancer.

of pinpointing the minority who have aggressive disease at the time of diagnosis. In standard diagnostic biopsies, pathologists examine up to a dozen biopsy cores and assign an overall grade called a Gleason score, based on changes in glandular architecture. At times, these scores do not match the verdict of post-surgical pathological analysis (which reveal actual pathology). Krasnitz and colleagues' new method augments these pre-surgical biopsy results with the power of single cell genomics. In collaboration with NYU and Cornell University medical centers, the team sequenced the genomes of several hundred single cells sampled from 8 patients' diagnostic biopsy cores. Using computational methods, the team looked for cells whose copy-number variation (CNV) profiles harbored the same irregularities, indicating possible cancerous tumors composed of clonal cells. Krasnitz and colleagues found their method was able to more accurately assess the tumors and more closely matched the post-surgical pathological analysis. Krasnitz believes single-cell analysis could significantly improve risk assessment and treatment decisions for prostate cancer, especially in cases of borderline Gleason scores.

Alexander J et al, “Utility of single cell genomics in diagnostic evaluation of prostate cancer” was published in *Cancer Research* January 15, 2018

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.

Predicting how splicing errors impact disease risk

In one collaborative project, Assistant Professor Justin Kinney has teamed up with another faculty member at CSHL, Adrian Krainer to predict how specific genetic mutations impact RNA splicing and thereby affect a person's risk for disease. Mutations at the first and second position of the 5' splice site (5'ss) are known to have a very strong impact while mutations elsewhere in the intron can have dramatic effects, no effect, or something in between, making it hard to predict how mutations at splice sites within disease-linked genes will impact patients. Mandy Wong, a postdoc shared by Kinney and Krainer through the **Interdisciplinary Scholars in Experimental and Quantitative Biology (ISEQB) Program** (see p. 7), led experiments using a massively parallel splicing assay (MPSA) in human cells to quantify the activity of over 32,000 5'ss sequences in three different gene contexts. The team found that, although splicing efficiency is mostly governed by the 5'ss sequence, there are substantial differences in this efficiency across gene contexts. These MPSA measurements facilitate the prediction of 5'ss sequence variants that are likely to cause aberrant splicing. This approach provides a framework to assess potential pathogenic variants in the human genome and streamline the development of splicing-corrective therapies. In future work, Kinney, Krainer, and Wong plan to apply a new method developed by CSHL Assistant Professor, David McCandlish and colleagues to understand patterns of genetic interaction in 5' splice sites. This method for inferring models of global epistasis was published in *PNAS* this year.



Mandy Wong

Wong MS et al, "Quantitative activity profile and context dependence of all human 5' splice sites" was published in *Molecular Cell* September 20, 2018



OTHER AFFILIATIONS AND ORGANIZATIONS

New York Genome Center

Several CSHL faculty members have affiliations with the New York Genome Center (NYGC), including SCQB member and Associate Professor Ivan Iossifov. This year, Adam Siepel also became an affiliate member of the NYGC and Molly Hammell joined the NYGC Center ALS Consortium. In addition, Siepel is co-chairing a new NYGC Working Group on Population Genomics, together with Dr. Eimear Kenny of the Icahn School of Medicine at Mount Sinai.

Meetings and Conferences

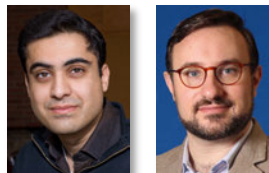
SCQB faculty members have co-organized the following meetings and conferences in 2018.

- **Probabilistic Modeling in Genomics (ProbGen)**, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, November 2018, Co-organized by Adam Siepel
- **Biological Data Science**, Cold Spring Harbor Laboratory, November 2018, Co-organized by associated member Michael Schatz
- **NY Area Population Genomics Workshop**, Cold Spring Harbor Laboratory, Cold Spring Harbor, January 2018, Co-organized by Adam Siepel

EDUCATION AND OUTREACH

THE SCQB SERVES AS A HUB for education, training and research in the quantitative sciences. Students, postdocs, and staff are encouraged to participate in regular symposia, weekly informal gatherings to discuss their research, and journal clubs which explore topics in deep learning and sequence-function relationships, as well as the genomics of gene regulation.

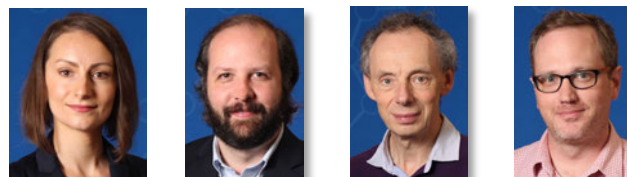
PhD training program



Mickey Atwal & Justin Kinney




We continue to offer a broad training program in quantitative biology to PhD students through the Watson School of Biological Sciences. In 2018, Assistant Professor Justin Kinney and Associate Professor Mickey Atwal led a 2.5-day **QB “Bootcamp”** which provides a rapid introduction to Python and the computer cluster at CSHL. This introduction is followed by a 16-week **Specialized Discipline Course in QB** with basic training in the following areas:

- Probability and Statistics (Atwal)
- Statistical Mechanics and Sequence-Function Relationships (Atwal and Kinney)
- Machine Learning (Atwal and Engel)
- Population Genetics (McCandlish)
- Evolution and Phylogenetics (Krasnitz)
- Genomics (Levy)
- Algorithms for Sequence Analysis (Levy and Siepel)



Tatiana Engel, David McCandlish, Alexander Krasnitz and Dan Levy

2018 DOCTORAL RECIPIENTS

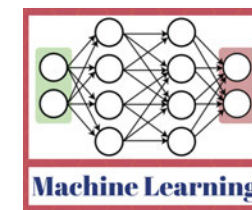
	Advisor	Academic Affiliation	Thesis
 Noah Dukler	Adam Siepel	Tri-Institute PhD Program in Computational Biology and Medicine	Statistical models for the function and evolution and cis-regulatory elements in mammals
 Talitha Forcier	Justin Kinney	Watson School of Biological Sciences	Precision measurement of cooperative interactions in transcriptional regulation in living cells
 Brad Gulko	Adam Siepel	Cornell University	Joint inference of human genomic function and selective pressure

Opportunities for postdoctoral researchers

The **Interdisciplinary Scholars in Experimental and Quantitative Biology Program (ISEQB)** was launched at the start of 2018. This innovative funding opportunity for postdoctoral research is designed to help recruit new postdocs or fund existing CSHL postdocs who are interested in cross-training in “wet” (experimental) and “dry” (computational) research laboratories. The program is aimed at both catalyzing collaborative research and promoting growth of the QB community at CSHL. This year’s scholars include: Wei Chia Chen working with Justin Kinney and Robert Maki, James Roach working with Tatiana Engel and Anne Churchland, and Molly Wong working with Justin Kinney and Adrian Krainer.

Advanced coursework in quantitative biology

This year, the SCQB began providing advanced coursework in quantitative biology to graduate students, postdocs and scientific staff. With support from the Watson School for Biological Sciences, the SCQB is offering a tutored version of Coursera’s **Machine Learning through a Massive Open Online Course (MOOC)** with onsite teaching assistance provided by SCQB postdoctoral researchers.



AWARDS AND RECOGNITION

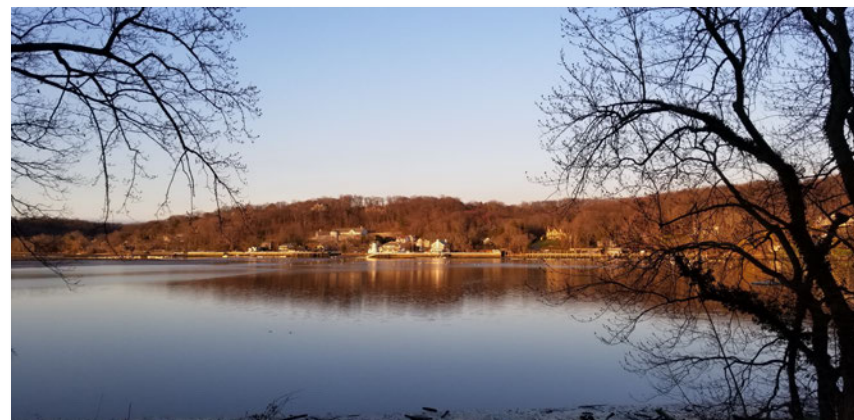
Molly Hammell awarded early career award



CZI Co-founder Priscilla Chan (left) speaking with Associate Professor Molly Hammell (right) at 13th annual Double Helix Medals dinner.

Associate Professor Molly Hammell was recently awarded the Chan Zuckerberg Initiative (CZI) Ben Barres Early Career Award for her proposed work on amyotrophic lateral sclerosis, better known as ALS or Lou Gehrig's disease. ALS is a neurodegenerative disease that causes complete paralysis and eventual death due to the rapid and progressive loss of motor neurons. Hammell proposes to develop machine-learning software that will systematically identify genetic factors and molecular

mechanisms that lead to motor neuron death. She will focus on transposable elements – viral-like genomic parasites that normally lay dormant in the genome – that are implicated in multiple diseases, including ALS. The award is part of CZI's Neurodegeneration Challenge Network, which connects researchers who are studying neurodegenerative diseases, and encourages a cross-disease perspective. The Ben Barres Early Career Acceleration Award focuses on investigators who are new to the field of neurodegeneration. Hammell is one of 17 researchers to win the award this year.



Cold Spring Harbor, NY (Photo courtesy: Yixin Zhou)

Tatiana Engel wins BRAIN grant



Tatiana Engel

Assistant Professor and associated member of the SCQB, Tatiana Engel, is developing computational tools for data collected specifically from the brain. Engel was recently awarded a Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative grant from the National Institutes of Health. The goal of the federal government's BRAIN Initiative is to accelerate the development and application of innovative technologies to produce a revolutionary new dynamic picture of the brain. Engel

will build mathematical models of decision-making activity in two different areas of the brain. She believes this initiative will play an important role in promoting theory and computation as an integral part of brain research. Working with two colleagues, CSHL Associate Professor Anne Churchland and Stanford University Professor Krishna Shenoy, Engel will create computational models based on their experimental data.

POSTDOCTORAL SPOTLIGHT



Yi-Fei Huang, PhD

Dr. Yi-Fei Huang joined Adam Siepel's lab in 2015 after completing his postdoctoral work at McMaster University in Ontario Canada. He has a background in evolutionary and statistical genetics with an interest in machine learning. While working on a wide variety of projects in the Siepel Laboratory, including predicting fitness consequences in the human genome, he also served as a tutor for the Specialized Course in QB in the Watson School. Dr. Huang will be starting as an Assistant Professor in the Department of Biology and the Huck Institute of the Life Sciences at Pennsylvania State University in January 2019.



**Cold
Spring
Harbor
Laboratory**