# ISCB NEWSLETTER

SPRING 2018 ISSUE 001

ISMB 2018 CALL FOR LATE POSTERS
Have you recently completed exciting research or are in the process? Want to Share? Submit a Late Poster to ISMB 2018!
Deadline: May 1, 2018

2018 ART in SCIENCE COMPETITION
Submit your image to the ISMB 2018 Art in Science Competition - The winning image will be featured in the next ISCB Newsletter!
Deadline: June 5, 2018

## ISMB 2018 ARRIVING SOON IN CHICAGO

ISCB's flagship meeting, Intelligent Systems for Molecular Biology (ISMB) 2018 is scheduled for July 6-10, 2018, at the Hyatt Regency Downtown, Chicago, Illinois. The conference is designed to make it easy for researchers sharing common interests to come together and listen to exciting new developments in their field. At the heart of the meeting are seventeen established communities (COSIs – Communities of Special Interest) reflecting most of the major research themes and training in computational biology, as well as research and novel techniques in emerging areas of computational biology, including intersections with other fields, which is presented in the General Computational Biology track.

Each day of the four-day conference includes distinguished ISMB Keynote Speakers (including the 2018 ISCB Scientific Award winners), COSI tracks, Special Sessions presenting emerging research areas that provide a more in-depth look at the subject matter, Poster sessions and a Technology Track featuring some of the latest tools, methods, and services for computational biology research. Within the COSI and General Computational Biology tracks you will hear 65 selected proceedings papers, highlight talks, and late-breaking research talks, as well as invited thought-leaders who are experts in the field and panel discussions. Delegates of the meeting will have the opportunity to participate in tutorial workshops, the ISCB Student Council Symposium, and Birds of a Feather sessions, as well as network with colleagues, friends, and potential collaborators in the Exhibition and Poster Hall while viewing posters and exhibits.

## IN THIS ISSUE



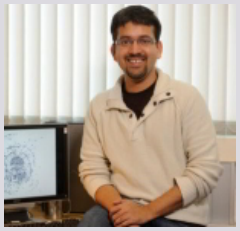
ISCB
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
www.iscb.org

## DISTINGUISHED KEYNOTE PRESENTATIONS

Steven Salzberg
Bloomberg Distinguished Professor of Biomedical Engineering, Computer Science, and Biostatistics Director;
Center for Computational Biology McKusick-Nathans Institute of Genetic Medicine;
Johns Hopkins University
Baltimore, United States

ISCB Overton Prize Award Keynote
Cole Trapnell
Assistant Professor, Department of Genome Sciences, University of Washington, United States

Martha L. Bulyk
Division of Genetics, Department of Medicine;
Department of Pathology;
Brigham & Women's Hospital and Harvard Medical School
Boston, United States



ISCB Innovator Award Keynote
M. Madan Babu
Programme Leader, MRC Laboratory of Molecular Biology,
Cambridge, United Kingdom

ISCB Accomplishments by a Senior Scientist Award Keynote
Ruth Nussinov
Senior Principal Investigator, National Cancer Instit ute, National Institutes of Health, United States;
Professor, School of Medicine, Department of Human Genetics, Tel Aviv University, Israel

## SAVE THE DATE!

**ISMB 2018**
July 6 – 10, 2018
Chicago, Illinois

**ISCB-LA SOIBIO EMBNET 2018**
Nov 5 – 9, 2018
Viña del Mar, Chile

**ROCKY 2018**
December 6 - 8, 2018
Aspen/Snowmass
Colorado

**RECOMB/ISCB Conference on Regulatory and Systems Genomics with DREAM**
December 8 - 10, 2018
New York City, USA

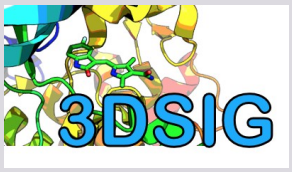
**GLBIO 2019**
May 19-22, 2019
Madison, USA

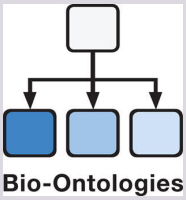
# ISCB COMMUNITIES OF SPECIAL INTEREST (COSI) TRACKS

The organized community sessions (COSI tracks) includes area specific keynote presentations, a selection of talks, which are featured in OUP Bioinformatics in the ISMB 2018 Proceedings supplement, as well as highlight and late-breaking research talks. The 2018 COSI Tracks feature the following communities of special interest:

## 3DSIG: Structural Bioinformatics and Computational Biophysics

It is impossible to fully understand biological systems without understanding the 3D structure of their constituting parts and their interactions. 3Dsig focuses on structural bioinformatics and computational biophysics and has become the largest meeting in this growing field.



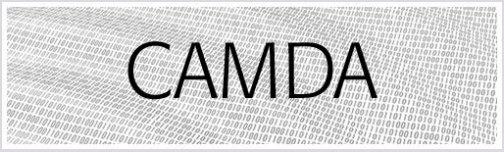




## Bio-Ontologies

Bio-Ontologies Community of Special Interest Group (COSI) covers the latest and most innovative research in the application of ontologies and more generally the organization, presentation and dissemination of knowledge in biomedicine and the life sciences.

## BioVis: Biological Data Visualization

The BioVis track aims to educate, inspire, and engage bioinformatics and biology researchers in state-of-the-art visualization research and visualization researchers in problems in biological data visualization.
Keynote speaker:
Martin Krzywinski, Staff Scientist, Bioinformatics, Genome Sciences Centre


## CAMDA: Critical Assessment of Massive Data Analysis

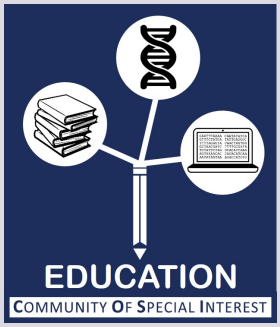
The large, complex data sets for the Critical Assessment of Massive Data Analysis
(CAMDA) contest include built-in truths for calibration. In an open-ended competition, however, both seasoned researchers and cunning students push the boundaries of our field, with unexpected questions or angles of approach often bringing the most impressive advances.

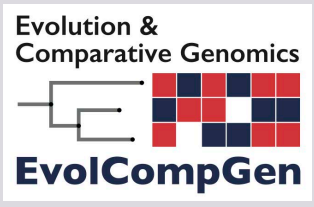
## Education: Computational Biology Education

Education-COSI focuses on bioinformatics and computational biology education and training across the life sciences with a goal to foster a mutually supportive, collaborative community in which bio scientists can share bioinformatics education and training resources and experiences, and facilitate the development of education programs, courses, curricula, etc., and teaching tools and methods.

Keynote speakers:
Russ Altman, Stanford University, United States, The Stanford Biomedical Informatics Curriculum: Early results from use in qualifying exams
Phillip Compeau, Carnegie Mellon University, United States, Establishing a computational biology flipped classroom
Anne Rosenwald, Georgetown University, United States, Bioinformatics in the Undergraduate Classroom: Barriers to Integration

Other invited speakers:
Bill Morgan, The College of Wooster, Ohio, United States, Development of the NIBLSE Learning Resource Collection and Incubators Allegra Via, Sapienza University of Rome, Italy, Cognitive psychology in the bioinformatics learning enterprise




## Evolution and Comparative Genomics

Evolution and comparative genomics are deeply intertwined with computational biology. Computational evolutionary methods, such as phylogenetic inference methods or multiple sequence alignment are widely used, yet remain far from “solved” and are indeed intense areas of research.

## Function SIG: Gene and Protein Function Annotation

The mission of the Function Community of Special Interest (Function-COSI) is to bring together computational biologists, experimental biologists, biocurators, and others who are dealing with the important problem of gene and gene product function prediction, to share ideas and create collaborations. The Function COSI features the Critical Assessment of Function Annotation, an ongoing community challenge aimed at improving methods for protein function prediction.
Keynote speaker:
Kimberly Reynolds, University of Texas Southwestern Medical Center, United States
Follow the developing program for Function SIG at http://biofunctionprediction.org/meetings/.

## HitSeq: High-throughput Sequencing

HiTSeq is a community of special interest devoted to the latest advances in computational techniques for the analysis of high-throughput sequencing (HTS) data. Sessions will be devoted to discussing the latest advances in computational techniques for the analysis of high-throughput sequencing (HTS) datasets and will provide a forum for in-depth presentations of the methods and discussions among the academic and industry scientists working in this field.

HiTSeq is honored to host as keynote speakers, thought leaders in the hot areas of HTS applications.
S. Cenk Sahinalp, Indiana University, United States
Ekta Khurana, Weill Cornell Medicine, United States
Nancy J. Cox, Vanderbilt University, United States

For developing information about the HiTSeq program, visit http://hitseq.org/.

## Junior Principal Investigator (JPI)

Transitioning from a post-doc to a junior PI can be a challenging process requiring careful planning. Once running a group, junior PIs are faced with many new tasks, some of which are learnt on the job. The Junior Principal Investigators group (JPI) aims to provide support during this process via a community of peers.

Birds of a Feather Sessions:
JPI Career development session: On leadership and management: focus on mentoring
July 7th 12:45-1:45
Panelists:
Trey Ideker, University of California San Diego, United States
Casey Green, University of Pennsylvania, United States
Lucia Peixoto, Washington State University, United states

Bioverly
The new biOverlay service provides peer review and editorial perspective on preprints and other academic outputs that are of particularly broad interest. https://bioverlay.org
Casey Green, University of Pennsylvania, United States

## MICROBIOME

The MICROBIOME Community of Special Interest aims at the advancement and evaluation of computational methods in microbiome research, especially metaomic approaches. Based on the Critical Assessment of Metagenome Interpretation (CAMI), the COSI supplies users and developers with exhaustive quantitative data about the performance of methods in relevant scenarios.

Keynote speakers:
Murat Eren, University of Chicago, United States Curtis Huttenhower, Harvard School of Public Health, United States
Adam Philippy, National Human Genome Research Institute, United States

## MLCSB: Machine Learning in Computational and Systems Biology

Systems Biology and Machine Learning meet in the MLCSB COSI. The community is the place for researchers of these areas to exchange ideas, interact and collaborate.

## NetBio: Network Biology

As large scale, systems-level data are becoming increasingly available, modeling and analyzing them as networks is widespread. Network Biology Community serves to introduce novel methods and tools, identify best practices and highlight the latest research in the growing and interdisciplinary field of network biology.

Keynote speakers:
Trey Ideker, UCSD, United States
Anna Goldenberg, University of Toronto, Canada

## RNA: Computational RNA Biology

RNA track covers the full range of research topics in the field of RNA Biology, from computational and high-throughput experimental methods development to their application in different aspects of RNA processing, structure, and function in both normal and disease conditions.

## RegSys: Regulatory and Systems Genomics

Regulatory genomics involves the study of the genomic control system, which determines how, when and where to activate the blueprint encoded in the genome. Regulatory genomics is the topic of much research activity worldwide. Since computational methods are important in the study of gene regulation, the RegSys COSI meeting focuses on bioinformatics for regulatory genomics.

## SysMod: Computational Modeling of Biological Systems

The Computational Modeling of Biological Systems (SysMod) aims to create a forum for systems modelers and bioinformaticians to discuss common research questions and methods. The session will focus on the conjoint use of mathematical modeling and bioinformatics to understand biological systems functions and dysfunctions.

Keynote speakers:
Andre Levchenko, Yale University, United States
Peter Sorger, Harvard Medical School, United States
John Tyson, Virginia Tech, United States

Following the developing program for the SysMod at https://sysmod.info/meetings/sysmod-2018/.

## TransMed: Translational Medical Informatics

TransMed covers the current developments in the field of clinical and translational medicine informatics. Analysis of large amounts of multi-omics, imaging (medical and molecular), mobile sensor, clinical and health records data is paving the way for precision medicine. In the TransMed track, we will explore the current status of computational biology and advance machine learning approaches within the field of clinical and translational medicine.

## VarI: Variant Interpretation

VarI-COSI track discusses the recent advances in the methodology for the annotation and analysis of genetic variants. The VarI-COSI meeting is dedicated to the recent advances in the analysis and interpretation of the genetic variants.

Keynote speaker:
Olga G. Troyanskaya, Princeton University, Simons Foundation, United States

# 14th ISCB Student Council Symposium

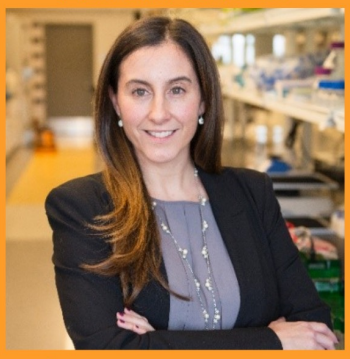
**Chicago, July 6, 2018**

## Keynote Speakers

**Debora Marks**

Department of Systems Biology
Harvard University

**Lucia Peixoto**



Washington State University

## Key Dates

~~**April 16**~~ **May 6:** Abstract Submission Deadline
**May 6:** Student Council Travel Grant Deadline
**May 28:** Abstract Acceptance Notification

**#SCS2018**
More information
symposium.iscbsc.org

facebook.com/iscbsc

twitter.com/iscbsc

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

ISCB Student COUNCIL
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

# PRE-CONFERENCE TUTORIAL WORKSHOPS

ISMB 2018 features half-day tutorial sessions on Friday, July 6, 2018, one day prior to the start of conference scientific program.

Delegates interested in the tutorials may register using the on-line registration system. Tutorial participants must be registered for the ISMB conference to attend a tutorial.

## Computational methods for comparative regulatory genomics

Changes in gene regulation have been repeatedly shown to contribute to phenotypic divergence. Until recently, most of our understanding of gene regulatory divergence on phenotypic complexity has been from a handful of examples. With the availability of a large number of sequenced genomes as well as functional genomics data including transcriptomic, epigenomic, and 3D genome organization across many species, we have the tremendous opportunity to comprehensively study how the gene regulation machinery changes globally across species and how this relates to global phenotypes. The availability of such multi-species datasets has fueled research in the development of systematic tools to perform comparative analysis of gene regulation at the sequence, expression, chromatin and network level. In this tutorial, we will review key challenges that arise in the comparative analysis of complex eukaryotes and cover recent computational algorithms and software tools that tackle some of these challenges. The tutorial will include invited talks from domain experts and will showcase theory and applications of tools to perform comparative analysis of molecular data at the sequence, expression, chromatin, and network level.

Specific topics include:
1 Identification and comparative analysis of regulatory sequence elements
2 Whole genome alignment
3 Phylogenetic tree construction
4 Comparative analyses of chromatin state and 3D genome organization
5 Inference and comparative analysis of transcriptional regulatory networks"

## Deep learning for network biology

Networks are ubiquitous in biology where they encode connectivity patterns at all scales of organization, from single cell to population level. Network approaches have been used many times to combine and amplify signals from individual genes, and have led to remarkable discoveries in biology, including drug discovery, protein function prediction, disease diagnosis, and precision medicine. Mathematical machinery that is central to these approaches is machine learning on networks. The main challenge in machine learning on networks is to find a way to extract information about interactions between nodes and to incorporate that information into a machine learning model. To extract information from networks, classic machine learning approaches often rely on summary statistics (e.g., degrees or clustering coefficients) or carefully engineered features to measure local neighborhood structures (e.g., network motifs). These classic approaches can be limited because these hand-engineered features are inflexible, they often do not generalize to networks derived from other organisms, tissues and experimental technologies, and can fail on datasets with low experimental coverage.

Recent years have seen a surge in approaches that automatically learn to encode network structure into low-dimensional representations using transformation techniques based on deep learning and nonlinear dimensionality reduction. The idea behind these representation-learning approaches is to learn a data transformation function that maps nodes to embeddings, points in a low dimensional space. Deep representation learning methods have revolutionized the state-of-the-art in network science. This tutorial will investigate methods and case studies for analyzing biological networks and extracting actionable insights, and in doing so; it will provide attendees with a toolbox of next-generation algorithms for network biology.

## High-throughput sequencing: Identification of disease variants in exomes and genomes

With the advent and the continuous drop in cost of next-generation sequencing, whole exome (WES) and whole genome sequencing (WGS) has become the platforms of choice for the diagnosis of Mendelian disease. A number of research studies are now generating WES or WGS data for sample sizes ranging from hundreds to thousands of cases. Identification of disease-causing variants, whether in clinical diagnostics or research studies, requires algorithms and statistical methods to score variants with respect to their likely relevance to the disease at hand. The goal of this tutorial is to present an overview of the current state of disease variant identification approaches with the more widely available whole-genomes and exome data of patients. We will discuss the most effective methods to interpret variants from patient's genomes for clinical diagnostics, as well as the statistical methods applied to the analysis of large cohorts for finding novel disease associated genes, considering rare variants.

At the end of the tutorial the participants will have an understanding of: 1) What are the challenges of analyzing WES/WGS data for clinical diagnostics and disease association studies; 2) How variant prioritization can be performed probabilistically and why its superior to empirical filtering schemes; 3) How to take advantage of family structures and phenotype information in these endeavors; 4) What are the difficulties in the analysis of rare variants for disease gene studies; 5) What are the typical and most advanced tools for rare variant analysis; and 6) What are the novel approaches for the analysis of disease cohorts for both identifying rare variants influencing common disease as well as ultra-rare homozygotes with very strong effects.

## Integrated Network Analysis: Cytoscape Automation using R and Python

Cytoscape is one of the most popular applications for network analysis and visualization. In this workshop, we will demonstrate new capabilities to integrate Cytoscape into programmatic workflows and pipelines using R and Python. We will begin with an overview of network biology themes and concepts, and then we will translate these into Cytoscape terms for practical applications. The bulk of the workshop will be a hands-on demonstration of accessing and controlling Cytoscape from R and Python to perform a network analysis of tumor expression and variant data.

An introduction to methods, analysis tools, and workflows for single cell RNA sequencing data

Single cell transcriptomics has emerged as a powerful tool to identify and interrogate novel cell types in homeostatic and perturbed states. Unlike bulk transcriptomics, single cell data provides resolution at the level of individual cells while working with much smaller quantities of RNA. As such, analysis of single cell RNA sequencing (scRNA-seq) data presents challenges of scale and technical noise, while providing the resolution necessary to pursue novel questions that earlier technologies did not allow.

The objective of the tutorial is to provide an overview of the laboratory and computational challenges involved in generating and analyzing scRNA-seq data. Participants will be introduced to popular molecular technologies for generating scRNA-seq data, and gain hands-on experience with existing software tools and computational methods for its analysis. The tutorial will briefly introduce approaches for preprocessing of scRNA-seq data, including demultiplexing, sequence alignment, and quality control. Then, starting from a cell x gene expression matrix, participants will learn standard methods to infer heterogeneity by identifying clusters of cells and perform analyses to assign cell identity and function. Participants will also be introduced to specialized analytical methods for exploring expression signatures of cell states, cellular differentiation trajectories, inference of cellular localization, and modern methods targeted towards better understanding of cancer biology. Analyses will be performed by executing commands in RStudio as well as leveraging newly developed point-and-click graphical R/Shiny interfaces.

### Machine learning methods in the analysis of genomic and clinical data

This tutorial covers various machine learning (ML) tools that have been developed for the analysis of genomic and clinical data. It is an intermediate level tutorial targeted to an audience with previous experience in diverse bioinformatics methods such as: i) genome-wide association studies, ii) comparison of structured data such as graphs or time-series, and iii) traditional text mining. State-of-the-art methods and their applications are presented. We will also discuss illustrative examples of how deep learning is currently being used in the analysis of biomedical data.

### Ontologies in Computational Biology

Ontologies have long provided a core foundation in the organization of biomedical entities, their attributes, and their relationships. With over 500 biomedical ontologies currently available there are a number of new and exciting new opportunities emerging in using ontologies for large-scale data sharing and data analysis. This tutorial will help you understand what ontologies are and how they are being used in computational biology and bioinformatics.

### Visualization of large biological data

The aim of this tutorial is to familiarize the participants with modern visual analytics methodologies applied to biological data and to provide simple hands-on training. Questions such as what is data visualization, what is visual analytics and how can biological data be visualized to gain insight are addressed, so that hypotheses can be generated or explored and further targeted analyses can be defined.

This course is designed for everyone who would like to learn and apply visualization techniques in the analysis of large biological data sets. The course provides useful background material on data visualization principles, but the focus is on methods and tools for visualization of next-generation sequencing data, other omics data and network data.

Topics covered are:

- Digital/Electronic Visualization of data
- Understanding color
- Visual Design Principles
- Examples of visualization of biological data
- Challenges of large-scale biological data visualization


The 2018 tutorial workshops are and more details can be found at URL:
**https://www.iscb.org/ismb2018-program/ismb2018-workshop**

# ISMB 2018 SPECIAL SESSIONS

ISMB 2018 Special Sessions track features emerging research areas of the field. Here are the 2018 Special Sessions.

**3D Genomics: Computational approaches for analyzing the role of three-dimensional chromatin organization in gene regulation.**
**Saturday July 7, 10:15am - 6:00pm**

Long-range gene regulatory interactions are defined as interactions between a region of regulatory DNA sequence and a target gene that can be hundreds of kilobases away. Such interactions are emerging as important determinants of cell type specific expression and the effect of regulatory sequence variants on complex phenotypes including those associated with diseases. The field of regulatory genomics has recently witnessed significantly increased interest in the three-dimensional structure of DNA in the nucleus, catalyzed by the availability of chromosome conformation capture (3C) data sets that characterize the 3D organization of chromatin at a genome-wide scale. This organization, also referred to as the 3D nucleome, is not only important for packing the genome into the nucleus, but it also has significant impact on how the genome functions. With the emergence of these new data types, there is an increasingly growing demand for computational tools that can systematically analyze these data. These tools range from data processing issues (e.g. mapping and normalization) to data analysis issues, such as predicting chromosomal organizational units (e.g. TADs), identifying significant interactions between regulatory elements (e.g. enhancer-promoter), examining the interplay of transcription factors, architectural proteins and chromatin states in establishing these interactions, and examining how these interactions are impacted by sequence variants.

**Single-particle Cryo-electron Microscopy, Cryo-electron Tomography, and Integrative/Hybrid Methods Studies of Macromolecular Machines: Opportunities and Challenges for the Bioinformatics Community**
**Sunday July 8, 10:15am - 12:30pm**

Among the most exciting of these newly deposited PDB structures are those coming from singleparticle cryo-electron microscopy (EM) and cryo-electron tomography (ET). Recent technical advances in sample preparation, electron optics, direct electron detection, and data processing software have created a perfect storm for the PDB. With these new methods cryo-EM and -ET are producing atomic level structures of macromolecular machines, such as multi-subunit RNA and DNA polymerases, ribosomes, and nuclear pore complexes.

The next wave of exciting new structures will come from so-called integrative/hybrid methods, which typically combine cryo-EM or -ET data with data from chemical cross-linking, fluorescence resonance energy transfer, and homology models to produce multi-scale structures of even larger biomolecular machines.

The Special Session will highlight examples of the exciting work going in in these two frontier areas of structural biology from four distinguished speakers, with reference to the manifold challenges and opportunities for the bioinformatics community.

**Omics Data Compression and Storage: Present and Future**
**Sunday July 8, 2:00pm - 6:00pm**

In 2003 the first human genome assembly was completed. It was the end of a project that took almost 13 years to complete and cost 3 billion dollars (around $1 per base pair). This milestone ushered in the genomics era, giving rise to personalized or precision medicine. Fortunately, sequencing cost has drastically decreased in recent years. While in 2004 the cost of sequencing a whole human genome was around $20 million, in 2008 it dropped to a million, and in 2017 to a mere $1000. As a result of this decrease in sequencing cost, as well as advancements in sequencing technology, massive amounts of genomic data are being generated. At the current rate of growth (sequencing data is doubling approximately every seven months), more than an exabyte of sequencing data per year will be produced, approaching the zettabytes by 2025. As an example, the sequencing data generated by the 1000 Genomes Project (www.1000genoms.org) in the first 6 months exceeded the sequence data accumulated during 21 years in the NCBI GenBank database.

In addition, the generation of other types of omics data are also experiencing a rapid growth. This situation calls for state-of-the-art, efficient compressed representations of massive biological datasets, that can not only alleviate the storage requirements, but also facilitate the exchange and dissemination of these data. This undertaking is of paramount importance, as the storage and acquisition of the data are becoming the major bottleneck, as evidenced by the recent flourishing of cloud-based solutions enabling processing the data directly on the cloud. For example, companies such as DNAnexus, GenoSpace, Genome Cloud, and Google Genomics, to name a few, offer solutions to perform genome analysis in the cloud.

This sentiment is also reflected by the NIH Big Data to Knowledge (BD2K) initiative launched in 2013, which acknowledged the need of developing innovative and transformative compression schemes to accelerate the integration of big data and data science into biomedical research. In addition, the International Standardization Organization (ISO) is developing, under MPEG (Moving Picture Expert Group), a standard for genomic information representation.

This special session will cover current efforts in this area, as well as future challenges. This is of importance to biologistics and researchers alike that work with omics data, as the developed tools will soon become part of their standard pipelines.

**Building an academia-industry bridge to bring precision medicine to the clinic**
**Championed by the ISCB Industry Advisory Council**
**Monday July 9, 2:00pm - 6:00pm**

Academia-industry collaboration is a principal vector for the translation of research discoveries from life sciences into pharma, the clinic and the public at large. The pharmaceutical, diagnostics, and healthcare technology industry have a long history of successful collaboration with academia, both in the form of clinical trials as well as basic, applied, and translational research. As a newer industry, biotechnology and bioinformatics companies have less history and often-different objectives than the pharmaceutical industry. As recent academic publications by companies such as Regeneron and 23andMe have demonstrated, academia-industry partnerships with strong bioinformatics focus can be well suited to add to the existing body of scientific knowledge.

Academia-industry partnerships exist in many forms such as consulting, contracted research, bilateral partnerships, public-private partnerships, start-ups, etc. These forms of collaboration invite several nuances that make project outcomes highly variable. This session will (1) highlight significant research produced by successful academia-industry partnerships that have resulted in a change in healthcare practices; (2) focus on the elements of the partnerships that enabled success in the context of value-driven bioinformatics enterprise and will brings together biologists, computer scientists, statisticians, medical doctors, startups, and industry professionals.

**SCANGEN: Single cell cancer genomics**
**Tuesday July 10, 10:15 pm - 4:30 pm**

In the past five years technological advances have given us the unprecedented ability to measure RNA and DNA at the single cell level. This now enables us to routinely measure gene expression and genomic alterations across tens of thousands of cells, discovering new cell types, developmental lineages, and cell-specific mutational patterns. This new data has prompted an explosion in statistical and computational methods development (http://
www.scrna-tools.org/) with over 150 tools being produced in the past few years.

To date the majority of methods developed have focused on either technical aspects (such as normalization and differential expression) or on applications in developmental biology such as lineage inference, with relatively little attention applied to the huge potential of single cell data to unveil the complex biology behind cancer inception and progression. As one of the first workshops of its kind, this special session will bring together researchers developing computational and statistical methods for single cell cancer biology. It will focus around
(though not be limited to) four core topics:

- Modelling cancer evolution
- Integrative analyses of multi-modal data
- Scalable inference at the single cell level
- Interactions and perturbations at the single cell level

**BD2K - Big Data to Knowledge**

This year, NIH BD2K is hosting a two-day track, July 7th-8th, in partnership with ISMB to highlight the accomplishments of the entire NIH BD2K Program as well as other data science activities at NIH. We have organized several special sessions on topics presenting state-of-the-art data science themes, including: i) Biomedical Data Science in Action; ii) BD2K Power Tools: Moving to the Cloud with Industrial Strength Data; iii) BD2K Data Visualization Tools & Future Directions; iv) Machine Learning Approaches to Enable Biomedical Discoveries; v) Building the FAIR Data Ecosystem for Discovery to Health; and vi) Young Investigator Award Competition. As NIH enters a new era of data sciences, this 2018 ISMB presents exciting opportunities for all investigators to connect with NIH Data Science Programs and to discuss future directions moving forward.

For more details about each session, **visit https://www.iscb.org/ismb2018-program/ismb2018-special-sessions**


# GCCBOSC 2018

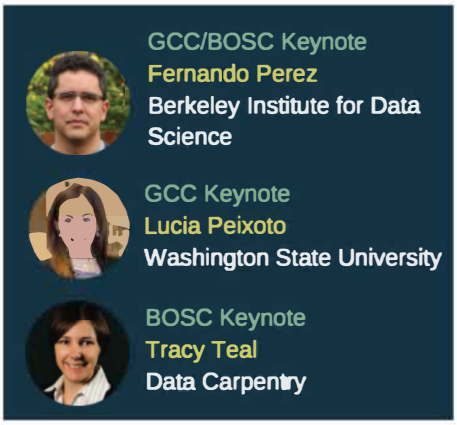
https://gccbosc2018.sched.com/

June 25-30, in Portland, Oregon, United States

The Bioinformatics Open Source Conference (BOSC) is organized by the Open Bioinformatics Foundation (OBF), an ISCB COSI. The OBF is a non-profit group dedicated to promoting the practice and philosophy of open source software development and open science within the biological research community.

Since 2000, BOSC has provided a forum for a wide range of developers and users to interact and share research results and ideas in open source bioinformatics and open science.

GCCBOSC 2018 will include two days of talks and posters, three keynote speakers, a panel discussion, Birds of a Feather, and more. The two days of talks will be preceded by two days of training, and will be followed by a two-day CollaborationFest.




## Early Registration for GCCBOSC ends May 11!

**Register now at** https://gccbosc2018.sched.com/

**Why is BOSC partnering with the Galaxy Community Conference (GCC) in 2018?**
In past years, BOSC has been part of the ISMB conference. Because of our continuing focus on broadening and deepening the BOSC community, we've been exploring ways to reach those in the bioinformatics community who aren't already part of the audience attracted by ISMB. After much discussion and planning, we decided to hold BOSC in conjunction with GCC in 2018. We would like to emphasize that this is an experiment for us and we are maintaining our connection as an ISCB COSI.

## The ISCB Student Council Internships Program: Calling All Principal Investigators

Emre Guney (Chair, Education and Internships Committee of ISCB-SC),
Farzana Rahman (Chair, ISCB Student Council)
Dan DeBlasio (ISCB Board of Directors, Student Council Representative)

The ISCB Student Council (ISCB-SC) is currently soliciting principal investigators (PIs) to host internships through the SC's Internships Program. The program connects students in developing nations with PIs in developed countries and helps students gain hands-on short-term research experience. These internships benefit the intern as well as the host lab both scientifically and culturally. The program launched in 2009 and is committed to providing computational biology training for students from developing nations and improving competencies in the computational biology field. The SC's education and internships committee (EIC) has helped coordinate eight internships over the past years.

The Internships Program plays an important role shaping the computational biology research worldwide and the feedback on the success of the program so far has been very positive from the students and PIs alike. The program relies on the PIs to host students from developing nations, and the SC strives to minimize the effort and time commitment from a PI for recruiting an intern by streamlining the entire pre-internship process, from advertising the position and collecting applications to screening the applications.

Through the Anna Tramontano Fellowship Fund, ISCB-SC is now able to not only advertise internship opening and screen the applicants but also provide funding to reduce the financial burden of hosting the students by contributing to their travel and visa acquisition expenses.

We are looking forward to receiving internship offers from PIs to increase the reach of the program and fortify its impact. Upon expression of interest by the PI, EIC facilitates finding a motivated and talented student whose interests match the host lab by advertising the position, evaluating applicants, and providing a shortlist of qualified candidates. We are reaching out to PIs from all over the world who could potentially host a student from developing nations in their lab.

If you're interested in hosting an internship or would like to know more about the Internships Program, you can read our recently published article in PLOS Computational Biology [1] and visit internships.iscbsc.org. If you have any questions, we welcome inquiries to internships@iscbsc.org!

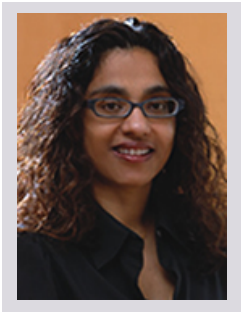
[1] http://dx.doi.org/10.1371/journal.pcbi.1005802

# ISCB Congratulates the 2018 Class of Fellows

**Terry Gaasterland, University of California San Diego,** for her service to ISCB since 1996 and her influence in the field of computational biology collecting and annotating pathways (originally through a tool called Magpie).

**Yves Moreau, Katholieke Universiteit Leuven R & D**, as one of the top-most formative leaders in computational biology in Europe. Having a background in engineering, he has made important and a broad range of contributions to the analysis of microarray data, to disease gene prioritization and to the analysis of genomics variants in the context of rare genetic diseases.

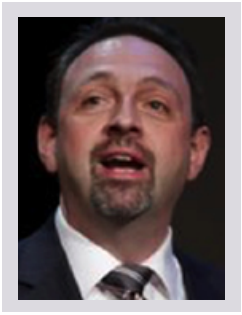
**William Pearson, University of Virginia,** in recognition for his development, distribution, and continuous improvements to FASTA and other similarity search methods, as well as his teaching of the biological and computational foundations of sequence analysis for more than 25 years.



**Mike Steel, Canterbury University, Christchurch New Zealand,** for outstanding contributions to mathematical and computational phylogenetics, and for service to the academic evolutionary biology research community.

**Patricia C. Babbitt, Univ. of California, San Francisco**, for her pioneering contributions to our understanding of sequence- structure-function connections in enzymes, and to our ability to computationally annotate and predict those connections.

**Hanah Margalit, Hebrew University**, as a research pioneer in the field, demonstrating excellence in teaching and research which led to many "firsts" in various subfields of bioinformatics, from structural biology to small-RNAs.



**Bernard M.E. Moret, Ecole Polytechnique Federale de Lausanne**, for mathematical theory and algorithms for comparative genomics and methods for understanding genome evolution, as well as his leadership in the computational biology research community.

**Mona Singh, Princeton University**, for being one of the first to bring machine learning and sophisticated algorithmic techniques to computational biology and making seminal contributions to a large number of the most important and topical challenges in the field, from protein folding to protein-protein interactions to systems biology.

**ISCB Response to NIH Proposed Data Science Strategic Plan**

To capitalize on the opportunities presented by advances in data science, the National Institutes of Health (NIH) is developing a *Strategic Plan for Data Science*. This plan describes NIH's overarching goals, strategic objectives, and implementation tactics for promoting the modernization of the NIH-funded biomedical data science ecosystem. As part of the planning process, NIH has published a draft of the strategic plan today, along with a Request for Information (RFI) to seek input from stakeholders, including members of the scientific community, academic institutions, the private sector, health professionals, professional societies, advocacy groups, patient communities, as well as other interested members of the public. ISCB's Public Affairs and Policy Committee submitted a response to the drafted strategic plan on behalf of ISCB.

The Committee found that the plan provides a framework to deal with data science challenges and recognizes the importance of data sciences to the overall success in meeting the NIH mission. Continuing the emphasis areas of Big Data to Knowledge (BD2K) initiative, it is highly positive that the Plan recognizes data standards, interoperability, infrastructure, and training as critical focus areas. In addition, the goals of defining different and appropriate funding mechanisms and reviewing criteria for data-science efforts will resolve some major impediments within the current NIH funding ecosystem.

However, as outlined in the suggestions below, the Proposed Strategic Plan needs to provide a more concrete and detailed roadmap in its rather unique treatment of: 1) distinction of databases and knowledge bases and their funding models; 2) separation of software tool development from repositories and curation efforts; and 3) a limited representation of the challenges of data use globally. Considering that databases, information portals, knowledge bases, and tool developers represent a significant portion of a highly heterogeneous scientific data ecosystem, imposing binary funding distinctions may lead to dysfunctional or simply unworkable solutions.

The conclusion (or view) of the ISCB is that the Proposed Strategic Plan is a much needed step in the right direction, but that before long-term funding policy decisions are made, there is a need for strong community input and buy-in to expand on the specifics and possible unforeseen consequences of vague definitions and distinctions as discussed in the changes suggested below.

### Eliminate Separation of Databases from Knowledge Bases

We recognize that the distinction between "databases" and "knowledgebases" is stressed in the Strategic Plan as a way of differentiating funding mechanisms. However, the wide variety of information portals renders such a binary distinction highly artificial. The Strategic Plan states that a DB makes available the "core data" (no definition is offered) of some biological system, whereas a KB organizes information "related to core datasets," and states that KBs typically require significant curation whereas DBs do not. Model organism DBs (MODs) are cited as an example of DBs, yet MODs have undergone hundreds of person-years of curation efforts – an apparent contradiction. MODs are Knowledgebases. Perhaps the confusion is between 'data stores' where data are not curated and there is minimal metadata assignments, and 'knowledgebases' where many relevant data stores are integrated and curated so as to facilitate the full use of the data for computational analysis.

This notion that "core data" is the key differentiating feature is not widely accepted among practitioners in the field and is quite vague. For example, the transcriptome is listed as core data (belonging to DBs), whereas an expression pattern is listed as belonging to KBs. But no clear separation into core versus related-to-core data is obvious in the following list of datatypes present in one well-known biological data/knowledge-base: genes, promoters, transcription factor binding sites, terminators,

operons, metabolites, enzymatic reactions, transport reactions, metabolic pathways, gene essentiality data. Since the Strategic Plan now seems to advocate that no efforts should be funded that combine both core data and non-core data, presumably the preceding data/knowledge-base project (and most MODs as well) must be divided into two separate projects. Such an element of the Plan would create a major obstacle for the information integration that is so valued by end users. In reality, a continuum exists between DBs and KBs, and attempts to find a reliable place on that continuum to define a funding policy would prove challenging and troublesome and may have to be abandoned.

**Proposed Separation of Software Tool Development**

We see the proposal to "Separate support for tools development from support for databases and knowledgebases" as quite problematic because in practice the developments of many tools (software) and DBs/KBs are tightly intertwined and their separation may be both impossible and ill advised. Often, tools developed independently from the participation of the DB/KB community fail to actively incorporate data and data updates and fall into dis-use.

For example, if a grant application is designed to develop the first DB for metabolomics data, it would be critical for this effort to develop software for parsing submitted data, validating submitted data, enabling curators to add and modify descriptions of the experimental conditions, storing submitted data to a database management system, and for powering a user website that allowed users to submit queries and view query results. Without such software, the DB could not be populated, checked for accuracy, nor made available to users. Without the software, there is no database! It is not clear whether the Strategic Plan means to imply that every DB/KB project must involve two grant applications, one for the software and one for the DB/KB, but we do not consider this to be an advisable process.

Another problem with this idea is the apparent underlying assumption that any third party can easily write software tools for a given database. This is not the case because every database has a schema – a precise computer definition of each type of data stored in the database (e.g., genes, proteins, metabolites), and the schema for a given database will change over time. Each software tool written for a given database must manipulate the data using exactly the same schema as the database currently uses, otherwise the database and the software will be incompatible.

We suspect that one motivation for the separation idea might be the notion that DB/KB grantees would include software development tasks to try to bolster the innovative appearance, and yet reviewers who appreciate the importance of the database might feel that the entire project needs to be scored highly, even when the proposed software is weak, to ensure the funding of an important database. Based on such assumptions, we suggest three alternative ways to view and solve this problem: (1) The assumption seems to be that DB/KB applicants frequently include poor quality tool development tasks in their proposals. It is likely that in most of the proposals the inclusion of high quality tool development tasks leads to high (better) scores because the proposed tasks are excellent. Do data exist regarding the frequency of high versus low-quality tool components in DB/KB applications? (2) If NIH develops improved review criteria for DB/KB applications that decrease the weighting and/or necessity of innovation, both grantees and reviewers will need to put less emphasis on innovation, which should largely solve the problem. (3) Particularly for large projects, reviewers should be encouraged to recommend excising project elements that they consider poor quality, to enable awarding of high scores to the remaining project elements. Imagine a DB application with an excellent operational plan but a weak plan for

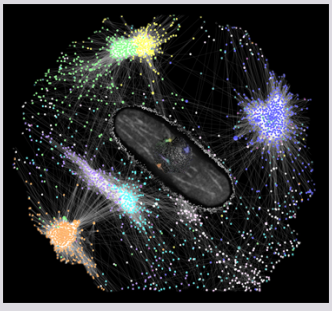
developing an innovative software tool. Everyone is served by continuing funding for the DB as a whole but not funding the software tool: users enjoy continued operation of the DB; the project retains highly skilled staff members. The grantees should be given an opportunity to resubmit the component for the software tool for later funding consideration. But: excision of project components must be performed judiciously.

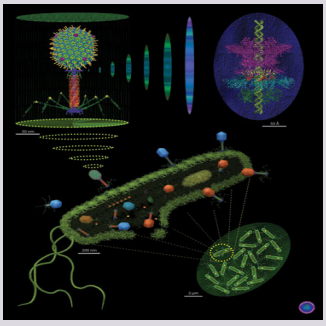
## ISCB Art in Science Competition --- Deadline: June 05, 20187

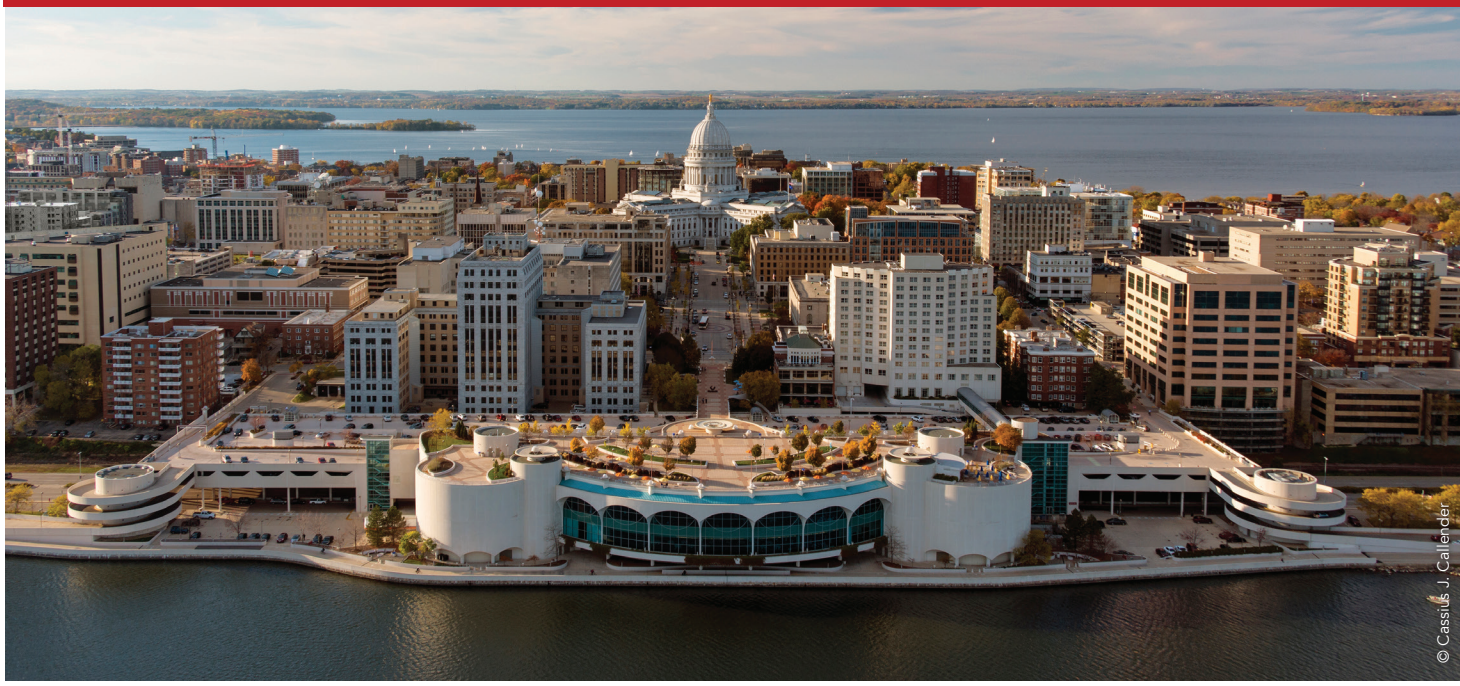
### ISCB invites submission to the 2018 ISCB Art in Science competition

All interested members (hereafter referred to as "artists") may submit images that have been generated as part of a research project and other creative efforts that involve scientific concepts or employ scientific tools and methods.

**https://www.iscb.org/iscb-art-in-science-competition**








# HAPPIEST CITY. HAPPIEST ATTENDEES.

It's no wonder National Geographic rated Madison, Wisconsin as one of the Top 10 Happiest Cities in the World. Madison boasts five sparkling lakes, and our State Capitol and the University Wisconsin-Madison campus are connected by a single, walkable street lined with local retail boutiques, restaurants and museums.

Enjoy meeting with colleagues at the Frank Lloyd Wright-designed Monona Terrace located in the heart of downtown Madison on the shores of Lake Monona. Connect to University of Wisconsin-Madison – home to state of the art research facilities and ranked #6 among U.S. universities in total research expenditures.

No matter the reason or the season, your visit to Madison is sure to be memorable.

**We look forward to welcoming attendees to these upcoming events:**

> Great Lakes Bioinformatics Conference – coming May 19–22, 2019

> Intelligent Systems for Molecular Biology, 2022 – coming July 8–15, 2022



## Don't miss out on a must-see international destination!

MADISON

GOING > BEYOND > VISIT™

GREATER MADISON CONVENTION & VISITORS BUREAU

VISITMADISON.COM

# UPCOMING AFFILIATED CONFERENCES

**DREAM Challenges and EPIDEMIUM@RECOMB**
Paris, France
Apr 19, 2018 through Apr 20, 2018
Event URL: http://recomb2018.fr/dream-challenges-and-epidemiumrecomb/
ISCB Member Discount: none

**RECOMB 2018**
Paris, France
Apr 21, 2018 through Apr 24, 2018
Event URL: http://recomb2018.fr/
ISCB Member Discount: 10 percent

**Chicago Genomics and Data Science Hackathon**
United States – Chicago, Illinois
Jun 11, 2018 through Jun 13, 2018
Event URL: http://ncbi-hackathons.github.io
ISCB Member Discount: none

**Bioinformatics Open Source Conference (BOSC)**
United States – Portland, Oregon
Jun 25, 2018 through Jun 30, 2018
Event URL: https://gccbosc2018.sched.com/
ISCB Member Discount: none

**ECCB 2018**
Athens, Greece
Sep 08, 2018 through Sep 12, 2018
Event URL: http://eccb18.org
ISCB Member Discount: 15 percent